STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:3581–3591
Published online 24 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2202

Multiple comparisons between two groups on multiple Bernoulli outcomes while accounting for covariates[‡]

James F. Troendle[*,†]

*Biometry and Mathematical Statistics Branch, Division of Epidemiology, Statistics, and Prevention Research,
National Institute of Child Health and Human Development, National Institutes of Health, DHHS,
Bld. 6100, Room 7B05, Bethesda, MD 20892, U.S.A.*

SUMMARY

The problem of adjusting for multiplicity when one has multiple outcome variables can be handled quite nicely by step-down permutation tests. More difficult is the problem when one wants an analysis of each outcome variable to be adjusted for some covariates and the outcome variables are Bernoulli. Special permutations can be used where the outcome vectors are permuted within each strata of the data defined by the levels of the (made discrete) covariates. This method is described and shown to control the familywise error rate at any prespecified level. The method is compared through simulation to a vector bootstrap approach, also using a step-down testing procedure. It is seen that the method using permutations within strata is superior to the vector bootstrap in terms of error control and power. The method is illustrated on a data set of 55 minor malformations of babies of diabetic and non-diabetic mothers. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: bootstrap; discrete; familywise error; logistic model; permutation

1. INTRODUCTION

Multiple comparison procedures that control the familywise error (FWE) rate for comparisons between two groups on each of multiple Bernoulli outcomes have been studied previously [1–3]. The most powerful method advocated in those papers is to perform a step-down test using resampling from the observed data vectors to estimate quantiles of the multivariate p-value distribution. Note that the observed data vectors are resampled ‘as is’. With continuous

[*] Correspondence to: James F. Troendle, DHHS/NIH/NICHD/DESPR/BMSB, National Institutes of Health, 6100 Executive Boulevard, Room 7B05, Bethesda, MD 20892, U.S.A.
[†] E-mail: jt3t@nih.gov
[‡] This article is a U.S. Government work and is in the public domain in the U.S.A.

Received 25 August 2004
Accepted 20 January 2005

data one would typically centre the data within each group (by subtracting the group mean from each value in the group) before resampling to better reflect the null hypothesis. With discrete data no centring is done so that the discreteness is preserved. Another important feature of the previous resampling methods with Bernoulli data is the lack of covariates. In this paper we shall study methods to test for group differences while accounting for covariates, and while controlling the FWE. These methods are new, and no other methods have been proposed for this purpose. Westfall and Young [2] describe methods for accounting for covariates in continuous regression models, but do not propose any such models for discrete outcome data. Although, resampling within blocks is mentioned in that book, it is intended there just to incorporate heteroscedasticity. In no way does resampling within blocks allow one to draw conclusions about regression coefficients. PROC MULTTEST [4] has no procedure appropriate for inference on a regression coefficient with a discrete outcome.

The motivating example for this paper is an analysis of multivariate Bernoulli data with a covariate. The data come from the Diabetes in Early Pregnancy Study (DIEP) [5]. The DIEP was an observational trial of diabetic and non-diabetic mothers followed during pregnancy until delivery. This analysis involves the presence or absence of 55 minor malformation types on each baby. Since investigators wanted to claim any observed significant difference represented a true association with diabetes, the testing should be performed with FWE control to limit spurious findings. Investigators also wanted to account for birthweight since it is known to be a potential confounder for many of the malformations.

Accounting for covariates with continuous outcome variables would typically be achieved by resampling from the residuals of a regression model. With discrete outcome measures, other approaches need to be considered. To measure association with Bernoulli outcomes one would fit a logistic regression model to each outcome with terms for each covariate and a term for group. The association of group status with the outcome is then measured by the coefficient of group status in the logistic regression model.

One resampling approach to multiplicity control is to bootstrap from the whole data vector which consists of outcomes, group indicator, and any desired covariates. Then, since the data have been generated by a process that does not necessarily follow the null hypothesis, the observed test statistic is compared to the difference between the test statistic from the bootstrap sample and the observed test statistic. This centres the set of bootstrap test statistics, but in general it is unclear how effective this approach will be in approximating the true null distribution.

If the covariates are discrete or can be approximated by discrete levels, there is another resampling option. For an analysis conditional on the covariate values, one can sample without replacement from the combined outcome vectors within each strata determined by the discrete covariate levels, assigning them in automatic fashion to the groups so that the number in each group equals the observed frequency within that strata. This type of 'resampling regardless of group' is that used by Westfall and Young [1, 2] and Troendle [3, 6], and has been compared to the usual bootstrap in the multi-dimensional case without covariates by Troendle *et al.* [7].

In this paper, we compare the multiplicity adjustment from the vector bootstrap to that from permuting within strata. The data model for multivariate Bernoulli outcomes is given in Section 2. In Section 3, the resampling methods are described and the FWE control for the permutation within strata method is proven. In Section 4, the methods are illustrated on the DIEP data on minor malformations. Section 5 contains a comparison of statistical properties by simulation. Extensions and recommendations are given in Section 6.

2. DATA MODEL

Assume there are k dichotomous outcome variables which will be represented by multivariate Bernoulli (MVB) vectors $\mathbf{Y}_i$ for $i=1,\dots,n$:

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{pmatrix} \sim \text{MVB}\left[\boldsymbol{\theta}_i = \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \vdots \\ \theta_{ik} \end{pmatrix}, \quad \boldsymbol{\Sigma}\right]$$

where $\boldsymbol{\Sigma}$ is an arbitrary covariance matrix. Here, i represents the subject index. The subjects are in either of two treatment groups, so we also have a treatment indicator $T_i=0$ or 1 for $i=1,\dots,n$. Finally, each subject is assumed to have p measured covariate values, denoted

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}, \quad i=1,\dots,n$$

The jth outcome is assumed to follow the logistic model,

$$\log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = a + \gamma_j T_i + \boldsymbol{\beta}_j' \mathbf{X}_i, \quad i=1,\dots,n \tag{1}$$

where

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j1} \\ \beta_{j2} \\ \vdots \\ \beta_{jp} \end{pmatrix}, \quad j=1,\dots,k$$

Interest lies in testing the null hypotheses $H_{0j}: \ \gamma_j = 0$, at level α, against a one- or two-sided alternative. We will confine attention to the two-sided case as the one-sided case is handled analogously. We desire to test each hypothesis H_{0j} while controlling the FWE, i.e. the probability of any type I error being committed.

Unadjusted p-values for testing H_{0j} are determined from $\hat{\gamma}_j$ and $\widehat{\text{var}(\hat{\gamma}_j)}$ obtained by maximum likelihood fitting of the logistic regression model (1):

$$P_j = 2\left[1 - \Phi\left(\left|\frac{\hat{\gamma}_j}{\sqrt{\widehat{\text{var}(\hat{\gamma}_j)}}}\right|\right)\right], \quad j=1,\dots,k$$

where $\Phi(x)$ denotes the cdf of a standard Gaussian random variable evaluated at x. Let

$$\mathbf{P} = \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_k \end{pmatrix}$$

3. RESAMPLING METHODS

3.1. Vector bootstrap

The usual bootstrap approach for regression settings (called bootstrapping pairs by Efron and Tibshirani [8]) is to consider the entire vector of observations on subject i:

$$\mathbf{Z}_i = \begin{pmatrix} \mathbf{Y}_i \\ T_i \\ \mathbf{X}_i \end{pmatrix}$$

A bootstrap data set, $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*$, is obtained by sampling with replacement n times from $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$. Notice the number of vectors, $\mathbf{Z}_i^*$, with $T_i = 0$ is random. We have used bootstrapping regardless of group, which seems most appropriate in consideration of the null hypothesis. If one wanted the number of vectors with $T_i = 0$ fixed, one might consider bootstrapping within group. Troendle *et al.* [7] showed in the multiple t-test case that when k is large relative to n, bootstrapping within group is exceedingly conservative but is otherwise similar to bootstrapping regardless of group. Another possibility is to bootstrap regardless of group and reassign the value of T_i randomly so that half of the vectors get $T_i = 0$. But this breaks the relationship between treatment and the covariates, and is not recommended. For this reason, we do not consider bootstrapping within group further. Calculate $\hat{\gamma}_j^*$ from the bootstrapped data as before, and calculate P_j^* from

$$P_j^* = 2\left[1 - \Phi\left(\left|\frac{\hat{\gamma}_j^* - \hat{\gamma}_j}{\sqrt{\widehat{\operatorname{var}(\hat{\gamma}_j^*)}}}\right|\right)\right], \quad j = 1, \ldots, k$$

Let

$$\mathbf{P}^* = \begin{pmatrix} P_1^* \\ P_2^* \\ \vdots \\ P_k^* \end{pmatrix}$$

Repeat this process B times yielding $\mathbf{P}^{*1}, \ldots, \mathbf{P}^{*B}$.

The vectors $\mathbf{P}^{*1}, \mathbf{P}^{*2}, \ldots, \mathbf{P}^{*B}$ can be used as a reference distribution for $\mathbf{P}$ in a step-down multiple comparison procedure as follows. Order the observed p-values

$$P_{(1)} \leqslant P_{(2)} \leqslant \cdots \leqslant P_{(k)}$$

and corresponding null hypotheses

$$H_{0(1)}, H_{0(2)}, \ldots, H_{0(k)}$$

Define the following subsets of $\{1, 2, \ldots, k\}$,

$$A_{(i)} = \{\text{indicies of } P_{(i)}, P_{(i+1)}, \ldots, P_{(k)}\}, \quad i = 1, \ldots, k$$

Step-down testing starts by testing $H_{0(1)}$ and proceeds sequentially until a null hypothesis is not rejected. At stage j, test $H_{0(j)}$ by rejecting if

$$\alpha_j^* = \frac{1 + \#\{b : \min_{r \in A_{(j)}}(P_r^{*b}) \leqslant P_{(j)}\}}{1 + B} \leqslant \alpha$$

If $H_{0(j)}$ is rejected, proceed by testing $H_{0(j+1)}$. Otherwise, stop testing.

Adjusted p-values for the method are obtained by enforcing monotonicity on the sequence α_j^*,

$$\alpha_j = \text{adjusted } p\text{-value for testing } H_{0(j)} = \max_{r \leqslant j} \alpha_r^*, \quad j = 1, \ldots, k$$

Because of experience with bootstrap and permutation procedures, it is conjectured that the vector bootstrap procedure asymptotically controls the FWE. However, no theory to support that conjecture is given here. The vector bootstrap is included because it is the natural extension of bootstrap methods to the multivariate setting.

3.2. *Permuting within strata*

Note that under the global null hypothesis, $\bigcap_{j=1}^k H_{0j}$, the random vectors $\{\mathbf{Y}_i : \mathbf{X}_i = x\}$ are exchangeable for each x. Thus, to obtain a null distribution, it is natural to permute the $\mathbf{Y}_i$ within each strata defined by the values of $\mathbf{X}_i$.

For concreteness, suppose there are C distinct values taken by $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$, and denote the $\mathbf{Y}_i$ vectors in the hth such strata, $\mathbf{Y}_{h,1}, \mathbf{Y}_{h,2}, \ldots, \mathbf{Y}_{h,n_h}$, where n_h is the number of subjects in strata h. Suppose there are $n_{h,0}$ subjects in strata h with $T_i = 0$. Consider the resampled data set, $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*$, obtained by sampling without replacement n_h times from $\{\mathbf{Y}_{h,1}, \mathbf{Y}_{h,2}, \ldots, \mathbf{Y}_{h,n_h}\}$ and assigning the first $n_{h,0}$ observations to have $T_i = 0$ and the remaining to have $T_i = 1$ ($\mathbf{X}_i$ values are held fixed at the strata values), independently for each $h = 1, \ldots, C$. Calculate $\hat{\gamma}_j^*$ from the permuted data as before, and calculate P_j^* from

$$P_j^* = 2\left[1 - \Phi\left(\left|\frac{\hat{\gamma}_j^*}{\sqrt{\widehat{\operatorname{var}}(\hat{\gamma}_j^*)}}\right|\right)\right], \quad j = 1, \ldots, k$$

Again, let

$$\mathbf{P}^* = \begin{pmatrix} P_1^* \\ P_2^* \\ \vdots \\ P_k^* \end{pmatrix}$$

Repeat this process B times yielding $\mathbf{P}^{*1},\ldots,\mathbf{P}^{*B}$.

The vectors $\mathbf{P}^{*1},\mathbf{P}^{*2},\ldots,\mathbf{P}^{*B}$ can be used as a reference distribution for $\mathbf{P}$ in an analogous manner to that given in Section 3.1 to produce a step-down multiple comparison procedure. However, it will be convenient to develop further notation here. Strict mathematical results pertain to the case where all possible permutations are considered, while in practice one would more likely be sampling independently for each strata from the full set of permutations within the strata (as described above). For any $D=\{d_1,d_2,\ldots,d_j\}\subseteq\{1,2,\ldots,k\}$, let y_D^{α} denote the largest number such that

$$\Pr{}^*\left\{\min_{i\in D}(P_i^*)<y_D^{\alpha}\right\}\leqslant\alpha$$

where $\Pr^*$ refers to the full multivariate permutation distribution, considering all permutations within each strata of the data (B distinct such permutations).

The step-down procedure starts by testing $H_{0(1)}$ and proceeds sequentially until a null hypothesis is not rejected. At stage j, test $H_{0(j)}$ by rejecting if

$$P_{(j)}<y_{A(j)}^{\alpha}$$

If $H_{0(j)}$ is rejected, proceed by testing $H_{0(j+1)}$. Otherwise, stop testing.

Adjusted p-values for the method are obtained by enforcing monotonicity on the sequence

$$\alpha_j^* = \frac{1+\#\{b:\min_{r\in A(j)}(P_r^{*b})\leqslant P_{(j)}\}}{1+B}$$

so that

$$\alpha_j = \text{adjusted } p\text{-value for testing } H_{0(j)} = \max_{r\leqslant j}\alpha_r^*, \quad j=1,\ldots,k$$

Proposition

Under the data model given in Section 2, the step-down procedure with independent permutations within each strata strongly controls the FWE at α.

Proof

The proof follows closely that of Korn *et al.* [9], if one considers their Procedure A with $u=0$ (no false discoveries allowed). Let K_0 be the set of indices corresponding to the variables that satisfy the null hypothesis. Let s be the rank of the p-value in the ordered list corresponding to the smallest p-value associated with a variable that satisfies the null hypothesis.

Then we have

$$\begin{aligned}\text{FWE} &= \Pr\{H_{0(1)}, H_{0(2)}, \ldots, H_{0(s)} \text{ are rejected}\}\\ &\leqslant \Pr\{P_{(s)} < y^{\alpha}_{A_{(s)}}\}\\ &\leqslant \Pr\{P_{(s)} < y^{\alpha}_{K_0}\}\\ &\leqslant \alpha\end{aligned}$$

The penultimate inequality is true because $y^{\alpha}_{A_{(s)}} \leqslant y^{\alpha}_{K_0}$ which follows because $K_0 \subseteq A_{(s)}$. Let $K_0 = \{d_1, d_2, \ldots, d_r\}$ and define $\mathbf{Y}_i^{(K_0)} \equiv (Y_{id_1} Y_{id_2} \cdots Y_{id_r})'$. Since $P_{(s)}$ is the smallest p-value corresponding to a variable with index in K_0, its distribution only involves p-values associated with variables satisfying the null hypothesis. Therefore, since the random subvectors $\{\mathbf{Y}_i^{(K_0)} : X_i = x\}$ are exchangeable for each x, the composite of the permutation distributions within each strata can be used as the reference distribution for $P_{(s)}$, verifying the last inequality. □

4. EXAMPLE

In this section, the methods described in Section 3 are applied to the DIEP study. In addition to the two methods described in Section 3, the results of applying the step-down Bonferroni method of Holm [10] is shown for comparison. This method controls the FWE strongly as long as the estimated logistic regression coefficients of (1) can be assumed to be mean zero Gaussian with independent chi-squared variance estimate under the null hypothesis. Since this is only true asymptotically, this procedure asymptotically controls the FWE strongly. The next section will compare the small and moderate sample size properties of the methods.

Briefly, this analysis is to determine if babies born to diabetic mothers have an increased risk of minor malformations. Diabetic ($n_1 = 467$) and non-diabetic ($n_2 = 277$) women were recruited in early pregnancy and followed until delivery, where the presence or absence of each of 55 minor malformations were noted. Each malformation was analysed in univariate fashion by Fisher's exact test, and the results of applying a resampling based step-down multiple test procedure is given in Reference [3] and summarized in Table I. Each null hypothesis is that the proportion of subjects with the malformation is the same in the two groups, whereas the alternative is that the proportion is higher in the diabetic group.

It is known that babies born to diabetic mothers weigh more than other babies, and weight could be a contributing factor in the diagnosis of certain malformations. Therefore, interest now is in determining if the regression coefficient for diabetic mother is significantly positive in a logistic regression model with treatment (diabetic mother) and birthweight as covariates. Birthweight was made discrete by using its observed quintiles, and assigning the scores from 1 to 5. The smallest six adjusted p-values obtained by applying the multiple testing methods ($B = 99\,999$) are shown in Table II. The results of the permutation within strata procedure is seen to be far less conservative than the procedure of Holm, while slightly less conservative than the vector bootstrap.

Table I. Unadjusted and adjusted p-values for DIEP data without correction for birthweight.

		Adjusted p-value
Malformation number	Unadjusted p-value	Step-down bootstrap
32	0.00033	0.0026
30	0.00097	0.0088
18	0.00916	0.1090
4	0.02424	0.2885
27	0.03290	0.3604
16	0.04228	0.4436

Table II. Unadjusted and adjusted p-values for DIEP data with correction for birthweight.

		Adjusted p-value		
Malformation number	Unadjusted p-value	Holm	Vector bootstrap	Within strata permutation
32	0.00028	0.0152	0.0009	0.0007
30	0.00243	0.1312	0.0151	0.0141
4	0.02839	1.0000	0.3119	0.3022
36	0.04043	1.0000	0.4499	0.4068
27	0.04501	1.0000	0.5010	0.4383
13	0.06332	1.0000	0.6815	0.6050

Comparing the multiplicity-adjusted p-values of Table I to the permutation within strata adjusted p-values of Table II, one sees that adjusting for birthweight has made a difference in the analysis. The multiplicity-adjusted p-value for malformation #30 was increased somewhat, while malformation #18 went from being third most significant to not making the top six.

5. SIMULATIONS

The statistical properties of the methods were evaluated by simulation. Equicorrelated (with correlation ρ) Bernoulli outcome variables were simulated with probability of event according to the logistic model (1) using program CBRAND of Ahn and Chen which is available at StatLib. The program uses truncation of multivariate Gaussian variates to generate correlated Bernoulli variates. The sample size in the two treatment groups were taken as equal. Each simulation had either one or two random covariates ($p=1$ or 2). Each covariate was independently generated with values 0 and 1. The first covariate took value 1 with probability 0.4 in the first treatment group and with probability 0.6 in the second treatment group. The second covariate (if included) took value 1 with probability 0.6 in the first treatment group and with probability 0.4 in the second treatment group. Each model had $a=-1.5$ and $\beta_{jl}=1.0$ for $j=1,\ldots,k$ and $l=1$ or $1,2$. The resampling methods used $B=199$ random

Table III. Simulated FWE with nominal level 0.05.

				Multiple comparison procedure		
k	n	p	ρ	Holm	Vector bootstrap	Within strata permutation
10	50	1	0.0	0.0064	0.0285	0.0496
			0.5	0.0057	0.0342	0.0475
			0.8	0.0036	0.0416	0.0413
	100	1	0.0	0.0262	0.0495	0.0444
			0.5	0.0218	0.0581	0.0504
			0.8	0.0128	0.0534	0.0449
		2	0.0	0.0321	0.0523	0.0491
			0.5	0.0267	0.0510	0.0486
			0.8	0.0165	0.0564	0.0519
50	50	1	0.0	0.0048	0.0074	0.0459
			0.5	0.0027	0.0139	0.0471
			0.8	0.0014	0.0348	0.0460
	100	1	0.0	0.0130	0.0514	0.0473
			0.5	0.0101	0.0562	0.0513
			0.8	0.0041	0.0581	0.0505
		2	0.0	0.0157	0.0517	0.0488
			0.5	0.0133	0.0516	0.0490
			0.8	0.0060	0.0589	0.0519
	500	1	0.0	0.0375	0.0470	0.0471
			0.5	0.0289	0.0529	0.0533
			0.8	0.0111	0.0518	0.0504
		2	0.0	0.0346	0.0476	0.0483
			0.5	0.0218	0.0472	0.0459
			0.8	0.0101	0.0500	0.0484

resamples since the FWE and power of the methods did not improve appreciably by increasing B beyond 199. All simulations used 10 000 replications.

Simulated FWE are shown in Table III for various correlation, sample size, and number of outcomes. The method of Holm is very conservative (although when there are a relatively large number of parameters compared to sample size the parametric p-values and therefore Holm's procedure can also be quite anti-conservative). The vector bootstrap is seen to be often more conservative than the permutation within strata method, while also being anti-conservative in some cases. The permutation within strata method appears to control the FWE as expected. Note that if the sample size per strata gets small (less than 25) the resampling methods become conservative due to the discrete resampling distribution. This is analagous to the extreme case of a permutation test based on two observations per group, for which one cannot obtain a p-value less than $\frac{1}{6}$. For this reason we restricted the results in Table III to cases where each strata had at least 15 observations.

To see that the less conservative nature of the permutation within strata method leads to higher power, data was simulated under an alternative with $\gamma_j = 1.5$ for each j. The power for

Table IV. Simulated average power with nominal level 0.05.

				Multiple comparison procedure		
k	n	p	ρ	Holm	Vector bootstrap	Within strata permutation
50	100	1	0.0	0.615	0.593	0.703
			0.5	0.629	0.661	0.745
			0.8	0.607	0.755	0.807
		2	0.0	0.540	0.512	0.629
			0.5	0.566	0.575	0.689
			0.8	0.554	0.706	0.771

each hypothesis was averaged. Results are given in Table IV. Notice that the permutation within strata method is the most powerful of those studied.

6. DISCUSSION

The permutation within strata method is useful in adjusting for multiplicity when testing each of multiple Bernoulli outcomes while correcting for the influence of discrete covariates. The method has been shown to control the FWE strongly, and simulations indicate it has higher power than other available procedures. Note that the proof of FWE control for the permutation within strata method only requires that the data vectors be exchangeable under the null hypothesis within each strata formed by the covariates. Thus, the method is valid quite generally for continuous or discrete-valued outcomes, although the advantage over alternative methods may be lost unless the outcomes are discrete. For example, the outcomes could be multivariate binary whose jth success parameter is linearly related as in (1).

If continuous covariates are to be accounted for, one may make discrete covariates from the continuous covariates by assigning scores to groups formed by appropriate sample quantiles. When doing so, one should be aware that the procedure may get very conservative if the number of subjects per strata is less than about 25. This is due to the highly discrete nature of the data which limits the set of possible distinct p-values. On the other hand, the usual tests based on univariate p-values from the estimated regression coefficients are not valid for such small sample sizes and furthermore may not be more powerful than the permutation within strata method. If the covariates can reasonably be made discrete while maintaining a sample size per strata of 25 or more, the permutation within strata method is recommended.

ACKNOWLEDGEMENTS

This study utilized the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD.

REFERENCES

1. Westfall PH, Young SS. *P* value adjustments for multiple test in multivariate binomial models. *Journal of the American Statistical Association* 1989; **84**:780–786.
2. Westfall PH, Young SS. *Resampling-based Multiple Testing*: *Examples and Methods for P-value Adjustment*, vol. 1. Wiley: New York, 1993.
3. Troendle JF. A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association* 1995; **90**:370–378.
4. SAS Institute Inc. *SAS/STAT Software*: *Changes and Enhancements through Release 6.12*, Cary, NC: SAS Institute Inc., 1997.
5. Mills JL, Knopp RH, Simpson JL, Jovanovic-Peterson L, Metzger BE, Holmes LB, Aarons JH, Brown Z, Reed GF, Bieber FR, Van Allen M, Holtzman I, Ober C, Peterson CM, Withiam MJ, Duckles A, Mueller-Heubach E, Polk BF, the NICHD-Diabetes in Early Pregnancy Study. Lack of relation of increased malformation rates in infants of diabetic mothers to glycemic control during organogenesis. *New England Journal of Medicine* 1988; **318**:671–676.
6. Troendle JF. A permutational step-up method of testing multiple outcomes. *Biometrics* 1996; **52**:846–859.
7. Troendle JF, Korn EL, McShane LM. An example of slow convergence of the bootstrap in high dimensions. *The American Statistician* 2004; **58**:25–29.
8. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993; 113.
9. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 2004; **124**:379–398.
10. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.