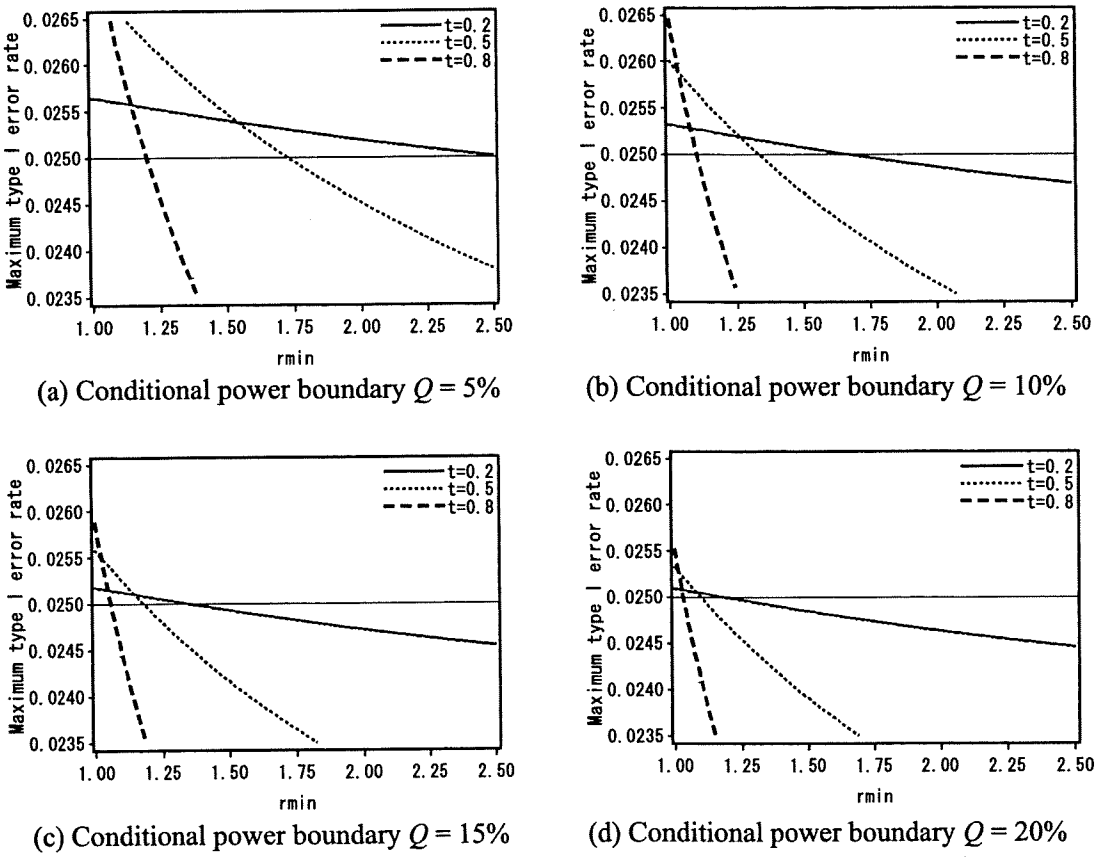


(a) Conditional power boundary $Q = 5\%$ (b) Conditional power boundary $Q = 10\%$

(c) Conditional power boundary $Q = 15\%$ (d) Conditional power boundary $Q = 20\%$

Fig. 1. Maximum type I error rate versus $r_{\min}$ (power $1 - \beta = 0.8$).

where $r_{\min}$ is a constant value larger than one. According to the rule (7), although one must accrue at least $(r_{\min} - 1)N_0$ subjects in addition to the initial planned N_0 ones, if one decides to increase sample size, one is allowed to set flexibly an upper limit of re-planned sample size N.

We will now consider the control of the maximum type I error rate under the rule (7). The maximum type I error rate $\alpha_{\max}$ is $\alpha + E\{I(r > r_{\min})\Delta_{\max}(r,t,z)\}$, where $\Delta_{\max}(r,t,z)$ is the maximum value of $\Delta(r,t,z)$ given t and z in $r > r_{\min}$ (see Appendix B). Figure 1 shows the plots of $\alpha_{\max}$ against $r_{\min}$ under the various scenarios. One can see that the maximum type I error rate $\alpha_{\max}$ decreases monotonically with the increase of the value of $r_{\min}$. Table 2 shows the minimum required value of $r_{\min}$ to control the maximum type I error rate under the nominal level, at which the line in Figure 1 crosses $\alpha_{\max} = 0.025$. When the conditional power boundary Q is set at 5 or 10 %, the minimum required value of $r_{\min}$ is about 2 at small information fractions, which means that the minimum required final sample size is $2N_0$ per group. In practice, the smaller value of $r_{\min}$ is set in the rule (7), the more flexibility will be obtained in the decision of final sample size. Therefore, we will set the conditional power boundary $Q = 20\%$ in the simulation studies.

Table 2. Minimum required value of $r_{\min}$ to control the maximum type I error rate under the nominal level of 2.5 %.

Conditonal power boundary Q (%)	Desired power $1-\beta$	Information fraction t								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	0.8	2.9	2.6	2.3	2.0	1.8	1.6	1.4	1.2	1.1
	0.9	1.9	1.8	1.7	1.5	1.4	1.3	1.3	1.2	1.1
10	0.8	1.9	1.7	1.6	1.5	1.4	1.3	1.2	1.1	1.1
	0.9	1.4	1.4	1.3	1.3	1.2	1.2	1.2	1.1	1.1
15	0.8	1.5	1.4	1.3	1.3	1.2	1.2	1.1	1.1	1.1
	0.9	1.3	1.2	1.2	1.2	1.1	1.1	1.1	1.1	1.1
20	0.8	1.3	1.2	1.2	1.2	1.1	1.1	1.1	1.1	1.1
	0.9	1.2	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1

3. Simulation Studies

3.1 Settings of simulations

We compared three methods of sample size re-estimation via simulation studies, which are the proposed modified CP approach, the original 50 %-CP one based on the rule (6) with $Q = 50$ and the weighted Z-statistic one based on the rule (2). In the simulations, we used the following rule as our proposed approach,

$$N = \begin{cases} N_0 & \textit{if} \quad CP(t,z,\delta=\hat{\delta}) < 20 \\ N_0 & \textit{if} \quad CP(t,z,\delta=\hat{\delta}) \geq 20 \quad \textit{and} \quad M \leq N_0 \\ r_{\min}N_0 & \textit{if} \quad CP(t,z,\delta=\hat{\delta}) \geq 20 \quad \textit{and} \quad N_0 < M \leq r_{\min}N_0 \\ M & \textit{if} \quad CP(t,z,\delta=\hat{\delta}) \geq 20 \quad \textit{and} \quad r_{\min}N_0 < M < N_{\max} \\ N_{\max} & \textit{if} \quad CP(t,z,\delta=\hat{\delta}) \geq 20 \quad \textit{and} \quad M \geq N_{\max} \end{cases}, \tag{8}$$

where $r_{\min} = 1.1$ or 1.2 according to Table 2.

Table 3 shows the settings of several design parameters. $\delta_{\min}$ and δ_{pre} denote a minimum clinically relevant and a pre-assumed effect size, respectively. α, β, $\delta_{\min}$, δ_{pre} and $N_{\max}$ were fixed values, and 18 scenarios of data were generated based on the combination of δ, t and N_0, that

Table 3. Settings of design parameters in simulations.

Nominal type I error rate α	Desired power $1-\beta$	Minimum relevant effect size $\delta_{\min}$	Pre-assumed effect size δ_{pre}	Maximum sample size $N_{\max}$	True effect size δ	Information fraction t	Initial sample size N_0
0.025	0.8	0.18	0.22	500	0.05 0.2 0.3	0.2 0.5 0.8	100 300

is, $3(\delta) \times 3(t) \times 2(N_0) = 18$. The three values of $\delta = 0.05$, 0.2 and 0.3 represent the situations for true effect size as nearly zero, moderate and large, respectively. Two values were considered as initial sample size N_0, where both of them represent the situations for smaller sample size than that calculated from the formula (1) under $\alpha = 0.025$, $\beta = 0.2$ and $\delta_{pre} = 0.22$, that is, $N_0 = 324$.

Assume that patients enter the trial with a uniform process at a fixed rate per unit time. An outcome variable Y_{ij} for subject i $(i = 1, \ldots, N_{\max})$ in group j $(j = 1, 2)$ was generated as normally distributed random variables with their mean, $\mu_1 = \delta$ and $\mu_2 = 0$, respectively, and common within variance $\sigma^2 = 1$. At an intermediate stage t, the estimate of effect size $\hat{\delta}$ and the conditional power defined as (4) were computed from the first n $(i = 1, \ldots, N_0 t)$ subjects per group accumulated so far. Re-planned sample size per group N was determined based on each rule (2), (6) and (8). Finally, the test statistic based on $2N$ subjects was computed from (3) and (5). The above process was replicated 10,000 times.

We compared three methods in terms of several operating characteristics such as power and average sample number (ASN). The overall and conditional evaluations were conducted, where the former is based on all cases including $N = N_0$, while the latter is based on only the cases of increasing sample size $(N > N_0)$. In the latter evaluation, several characteristics can be compared conditional on the decision of sample size adjustment.

3.2 Results

Table 4 shows the results. We presented only the case where $t = 0.5$ to save space, because similar results were obtained in other settings. In Table 4, Power(N_0) and Power(N) are powers based on fixed design of N_0 and re-planned N subjects, respectively, ASN is the average sample size, N_0^* is the required sample size for the fixed design to achieve the same power as Power(N), $\Pr(N > N_0)$ is the proportion of the cases of increasing sample size, and $\Pr(N = N_{\max})$ is the proportion of the cases of increasing sample size to the maximum number.

We first consider the case of $\delta = 0.2$ (moderate treatment effect). In the overall evaluations at $N_0 = 300$, the proposed method (20 %-CP) increased power 7 % compared with the fixed design, while its gain in the original one (50 %-CP) was 3 %. The weighted Z-statistic approach (weighted-Z) attained the highest power (0.84) and had the largest ASN (400), because the approach decided to increase sample size more frequently, that is, $\Pr(N > N_0) = 0.60$ and $\Pr(N = N_{\max}) = 0.42$. Two CP methods had nearly the same efficiency as the fixed design (the ratio $\mathrm{ASN}/N_0^* = 1.02$, 1.00, respectively), while the weighted-Z was a little more efficient $(\mathrm{ASN}/N_0^* = 0.93)$. In the case of small initial sample size $N_0 = 100$, the proposed 20 %-CP increased power 9 %, improving the gain of the original 50 %-CP (2 %), while both CP methods had the far below power from the desired power (0.8). Although the weighted-Z improved the power much more than CP methods owing to the much more frequent increase of sample size, it could not reach to the desired power and its efficiency was less than the fixed design.

Table 4. Simulation results ($t = 0.5$).

Methods	N_0	Overall evaluations							Conditional evaluations					
		Power(N_0)	Power(N)	Pr($N > N_0$)	Pr($N = N_{\max}$)	ASN	N_0^*	ASN/N_0^*	Power(N_0)	Power(N)	ASN	N_0^*	ASN/N_0^*	β error†
Effect size is moderate: $\delta = 0.2$														
50 %-CP	300	0.69	0.72	0.23	0.05	327	321	1.02	0.74	0.86	414	466	0.89	0.12
20 %-CP	300	0.69	0.76	0.38	0.20	356	358	1.00	0.67	0.87	448	472	0.95	0.13
Weighted-Z	300	0.69	0.84	0.60	0.42	400	432	0.93	0.53	0.77	466	368	1.27	0.27
50 %-CP	100	0.30	0.32	0.19	0.00	109	111	0.98	0.46	0.58	146	236	0.62	—
20 %-CP	100	0.30	0.39	0.34	0.00	139	141	0.98	0.37	0.64	212	271	0.78	—
Weighted-Z	100	0.30	0.66	0.83	0.42	325	284	1.14	0.20	0.65	371	273	1.36	0.34
Effect size is large: $\delta = 0.3$														
50 %-CP	300	0.96	0.96	0.15	0.03	316	313	1.01	0.93	0.98	405	360	1.12	0.003
20 %-CP	300	0.96	0.97	0.22	0.09	329	334	0.99	0.91	0.99	434	390	1.11	0.003
Weighted-Z	300	0.96	0.99	0.27	0.15	339	393	0.86	0.85	0.98	445	342	1.30	0.03
50 %-CP	100	0.57	0.60	0.22	0.00	110	109	1.01	0.66	0.80	144	176	0.82	—
20 %-CP	100	0.57	0.67	0.37	0.00	139	128	1.08	0.58	0.86	204	207	0.98	—
Weighted-Z	100	0.57	0.90	0.67	0.27	257	235	1.09	0.41	0.91	332	240	1.38	0.03
Effect size is nearly zero: $\delta = 0.05$														
50 %-CP	300	0.09	0.09	0.11	0.03	314	331	0.95	0.26	0.28	428	1539	0.28	0.76
20 %-CP	300	0.09	0.10	0.24	0.16	340	380	0.89	0.19	0.23	468	1197	0.39	0.81
Weighted-Z	300	0.09	0.11	0.93	0.84	476	449	1.06	0.06	0.09	490	280	1.75	0.94
50 %-CP	100	0.06	0.06	0.09	0.00	104	115	0.91	0.18	0.19	147	961	0.15	—
20 %-CP	100	0.06	0.06	0.20	0.00	127	153	0.83	0.13	0.17	235	804	0.29	—
Weighted-Z	100	0.06	0.08	0.94	0.60	395	249	1.58	0.04	0.06	414	149	2.77	0.97

† β error is the conditional type II error rate given $N = N_{\max}$, and the symbol "—" means the case where the conditional β error could not be computed because of $\Pr(N = N_{\max}) = 0$.

In the conditional evaluations at $N_0 = 300$, the power gain of the 20 %-CP compared with the fixed one was 20 % and it was higher than that of the 50 %-CP (12 %). These conditional Power(N) were higher than that of the weighted-Z, which means that the power of the weighted-Z given the increase of sample size was low, although it increased sample size more frequently. For the conditional ASN, both CP methods were more efficient than the fixed one, while the weighted-Z required additional about 100 subjects compared with the fixed one. Furthermore, the conditional β error given $N = N_{\max}$ was high in the weighted-Z, which means that there were much possibilities of not detecting the significant result in spite of increasing the sample size to its maximum. When $N_0 = 100$, the power gain of the 20 %-CP (27 %) was much larger than that of the original 50 %-CP (12 %). While there was no difference in the conditional Power(N) between the 20 %-CP and the weighted-Z, the conditional β error of the latter method was high.

Next, we consider the case of $\delta = 0.3$ (large treatment effect). As was expected, when $N_0 = 300$, the overall power of the fixed design was high and power gains in three re-estimation methods were little. There were no remarkable differences in the ASN between three methods. When $N_0 = 100$, overall power gains in the 50 %-CP, 20 %-CP and weighted-Z compared with the fixed design were 3 %, 10 % and 33 %, respectively, while there was no large difference in the conditional Power(N) between the 20 %-CP and the weighted-Z.

Finally, we consider the case of $\delta = 0.05$ (nearly zero treatment effect). The weighted-Z increased sample size erroneously in more than 90 % case and most of them were attained to its maximum (in result, ASN were nearly $N_{\max}$), while cases of sample size increase by both CP ones were fewer than those observed at $\delta = 0.2$. The ASN of the 50 %-CP was nearly the same as the initial sample size N_0, while a few increase was observed in the 20 %-CP.

4. Discussion

4.1 Implications from simulation results

In this paper, we proposed a new sample size re-estimation method based on the rule (7) and compared it with the original 50 %-CP and the weighted Z-statistic ones. Our proposed method can control the type I error rate flexibly due to the restriction on the minimum required sample size ratio $r_{\min}$. For example, $r_{\min} = 1.1$ implies that the type I error rate can be preserved under the nominal level if at least $1.1N_0$ subjects per group are accrued, even when re-planned sample size based on the rule (6) is $2N_0 = M < N_{\max}$. As shown in Table 2, when the conditional power boundary Q is set at 20 %, most of the values of $r_{\min}$ are 1.1 or 1.2 and thus, the requirement will not be so impractical that the final sample size can be determined flexibly accounting for circumstances of the trial concerned. Furthermore, we calculated the possible inflation of the maximum type I error rate under $r_{\min} = 1.0$ and $Q = 20$, and it was found to be 0.05 % in the worst case. That means the type I error rate would not be materially inflated even if the restriction was not kept at all. Thus, we recommend Q is set at 20 %. On the other hand, when Q

is set at 5 or 10 %, the values of $r_{\min}$ are more than 1.4 especially at small information fractions. In real clinical trials, since the minimum sample size increment will be influenced by not only predictable factors but also many uncertain ones, one may hesitate to make a decision based on an interim result at an early stage. In other words, it is not recommended to consider increasing sample size at a very early stage with little available information because of the imprecise estimate of effect size as well as the possible inflation of the type I error rate.

It was confirmed from the simulation results that the original 50 %-CP method is conservative, because the probability of increasing sample size $\Pr(N > N_0)$ was small with the result that the overall power was nearly the same as the fixed design. On the other hand, our proposed one increased power about 10 % compared with the fixed design by setting the value of Q to a lower one. Therefore, the proposed method is more useful for reducing the risk of not detecting the treatment effect of medical interest but slightly smaller than what was expected.

The conditional power of our method was higher than that of the weighted Z-statistic one, although the latter was superior in the overall power. In real clinical trials, additional costs must be paid for increasing sample size as well as conducting an interim look. Therefore, it is also important to improve the conditional power when actually there is a treatment effect. Using this high conditional power characteristic, the proposed method can be applied as follows. Consider a large prevention trial, in which a huge amount of fund will be needed to accrue a large number of healthy subjects in a fixed design with a power of 90 %. Because the existence of a clinically meaningful treatment difference is usually uncertain, researchers have to take a considerable economic risk and thus hesitate to start the trial. In this scenario, initial sample size calculation is conducted based on the minimum required power such as 70 % or 80 %, and if the conditional power based on the interim estimate of effect size is greater than 20 %, sample size re-estimation using the rule (7) is conducted to obtain the target power of 90 %. Such a strategy is useful for reducing not only the economic risk but also the conditional average sample number.

When considering sample size re-estimation, it is also important to control the error probability of increasing sample size under no treatment effect. The error probability of the weighted Z-statistic one was very high (93 % in $N_0 = 300$ and 94 % in $N_0 = 100$), while that of our method was small (24 % in $N_0 = 300$ and 20 % in $N_0 = 100$). Furthermore, the probabilities of attaining a maximum sample size were 84 % and 16 % ($N_0 = 300$), 60 % and 0 % ($N_0 = 100$), respectively. Increasing sample size when the interim result shows no treatment effect usually requires a dramatic sample size increase which may not be affordable in practice and waste limited resources.

Despite our method's superiorities over the weighted Z-statistic one in terms of the conditional power and probability of erroneously increasing sample size under no treatment effect, our method was inferior in overall power, especially for small initial sample size. The feature is due to the lower probability of increasing sample size under a moderate or large treatment effect than the weighted Z-statistic approach. Furthermore, because the conditional power $CP(t, z, \delta = \hat{\delta})$,

which is used for the decision whether or not to increase sample size, is based on the initial sample size, and the chance of increasing sample size will become less as the smaller sample size is planned. Thus, the conditional power boundaries in our method should be tuned on account of such trade-off between the high conditional power and low erroneously increasing probability, and the low overall power. If a trial sponsor overwhelmingly regards the overall power as importance, one should use the weighted Z-statistic approach.

In real clinical trials with a confirmatory purpose, it is usual to take a conservative approach so long as researchers do not have a firm prior information on treatment effect, and is rare to assume a large treatment effect. When actually there is a large treatment effect, initial sample size will be larger than true required one, and all sample size re-estimation methods including our proposed one will add the extra subjects. However, in this situation, if sample size re-estimation is considered in the context of a group sequential design, the formal interim analyses will stop the trial for efficacy in the presence of an overwhelming treatment difference, and thus, the above mentioned problem will not be much concerned.

4.2 Group sequential settings

As argued by some authors (Cui, Hung, and Wang, 1999; Shih, 2001), it may be more attractive if sample size re-estimation and efficacy interim analyses can be used in the same trial. The proposed method can be extended to a group sequential trial, where a decision may be made at the interim analysis to stop the trial early due to a convincing treatment benefit or to increase sample size if the observed result is not as good as expected. The extension must take account of not only the multiplicity of the hypothesis testing, but also the distributional change in the final test statistic due to the early stopping by interim analyses. Small limited simulation studies were conducted to evaluate the behaviors of our proposed method in the setting of a group sequential design. The relationships between $\alpha_{\max}$ and $r_{\min}$ were not different from Figure 1, and the values of $r_{\min}$ to control the maximum type I error rate under the nominal level were essentially the same as Table 2. However, in futility stopping settings, our simulation results might change on some aspects. For example, if a large futility stopping boundary is set (e.g. stop a trial for futility if $CP(t, z, \delta = \hat{\delta})$ is less than 10%), differences in conditionally evaluated characteristics between the CP and weighted-Z methods become moderate due to the conditional type II error rate reduction of the weighted-Z method. Differences in the error probability of increasing sample size under no treatment effect might also become moderate. When the required sample size is far above the maximum sample size, it is not ethical to continue a trial with sample size adjustment. Further research is needed to assess the characteristics of those re-estimation methods with various futility stopping criterion.

4.3 Some remarks

Although sample size re-estimation is an appealing design, it should not substitute for careful planning of a trial (Gallo, and Maurer, 2006; Schäfer, 2006; Wittes, and Lachenbruch, 2006). One

must note that both CP and weighted-Z approaches cannot improve the power to the desired level when initial sample size is much smaller than true required one. As described by many authors (Chen, DeMets, and Lan, 2004; Gould, 2001; Shih, 2001; Hung et al., 2006; Koch, 2006), important regulatory and logistical issues remain unresolved such as who will see what data; what knowledge of the trial result needs to be protected from investigators, patients and the sponsor's management; how to minimize possible influence of sample size re-estimation on investigators and patients behavior during the trial; whether or not the knowledge from external trials will have adverse influence on the current trial, etc. While many discussions on such issues have been conducted in the above referred papers, and the PhRMA Working Group on Adaptive Designs (Gallo et al., 2006) was formed in order to develop general consensus, the further research at individual trial level is still needed to investigate the impact of such a decision based on unblinded interim results on the conduct of the trial, to find ways to protect the integrity of the study, and to assess the risk benefit of such designs.

Acknowledgements

We would like to thank two referees for their many helpful comments and suggestions. We also thank one referee for providing the Mathematica program to calculate the actual or maximum type I error rate.

REFERENCES

Bauer, P., and Kohne, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**, 1029-1041.

Burman, C. F., and Sonesson, C. (2006). A flexible design is sound? *Bioimetrics* **62**, 664-683.

Chen, Y. H. J., DeMets, D. L., and Lan, K. K. G. (2004). Increasing sample size when the unblinded interim result is promising. *Statistics in Medicine* **23**, 1023-1038.

Chow, S-C., and Chang, M. (2007). *Adaptive Design Methods in Clinical Trials.* Chapman and Hall/CRC.

Cui, L., Hung, H. M. J., and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853-857.

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645-2660.

Fisher, L. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551-1562.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development-an executive summary of the PhRMA Working Group. *Jorunal of Biopharmaceutical Statistics* **16**, 275-283.

Gallo, P., and Maurer, W. (2006). Challenges in implementing adaptive designs: comments on the viewpoints expressed by regulatory statisticians. *Biometrical Journal* **48**, 591-597.

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* **11**, 55-66.

Gould, A. L. (2001). Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine* **20**, 2625-2643.

Hung, H. M. J., O'Neill, R. T., Wang, S. J., and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* **48**, 565-573.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical Journal* **48**, 574-585.

Lehmacher, W., and Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials. *Biometrics* **55**, 1286-1290.

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953-969.

Proschan, M., and Hunsberger, S. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315-1324.

Schäfer, H. (2006). Adaptive designs from the viewpoint of an academic biostatistician. *Biometrical Journal* **48**, 586-590.

Shen, Y., and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190-197.

Shih, W. J. (2001). Comment on Type I error in sample size re-estimations based on observed treatment difference (p. 497-513). *Statistics in Medicine* **20**, 515-518.

Shun, Z., Yuan, W., Brady, W. E., and Hsu, H. (2001). Type I error in sample size re-estimations based on observed treatment difference. *Statistics in Medicine* **20**, 497-513.

Wittes, J., and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65-72.

Wittes, J., and Lachenbruch, P. A. (2006). Opening the adaptive toolbox. *Biometrical Journal* **48**, 598-603.

Wittes, J., Schabenberger, O., Zucker, D., Britain, E., and Proschan, M. (1999). Internal pilot studies I: type I error rate of the naive t-test. *Statistics in Medicine* **18**, 3481-3491.

Zucker, D., Wittes, J., Schabenberger, O., and Brittain E. (1999). Internal pilot studies II: comparison of various procedures. *Statistics in Medicine* **18**, 3493-3509.

Appendix

A. Calculation of the actual type I error rate under the rule (6)

From the equation (4), the value of test statistic at an intermediate stage under the assumed CP boundary Q, $z_{CP=Q}$, can be written as $z_{CP=Q} = \sqrt{t}\left(\sqrt{1-t}\Phi^{-1}(Q/100) + z_\alpha\right)$. Let $z_{\lambda=1} = \sqrt{t}(z_\alpha + z_\beta)$ and $z_{\lambda^2=N_{\max}/N_0} = \sqrt{tN_0/N_{\max}}(z_\alpha + z_\beta)$, because there is a relationship between z and $\lambda = \delta_{pre}/\hat{\delta}$ as $z^2 = t(z_\alpha + z_\beta)^2/\lambda^2$. Using these notations, the rule (6) can be written in terms of the sample size ratio r as,

$$r = \begin{cases} 1 & if \quad z < z_{CP=Q} \\ 1 & if \quad z \geq z_{CP=Q} \quad and \quad z \geq z_{\lambda=1} \\ \lambda^2 & if \quad z \geq z_{CP=Q} \quad and \quad z_{\lambda^2=N_{\max}/N_0} < z < z_{\lambda=1} \\ N_{\max}/N_0 & if \quad z \geq z_{CP=Q} \quad and \quad z \leq z_{\lambda^2=N_{\max}/N_0} \end{cases}.$$

$z_{CP=Q}$ is always less than $z_{\lambda^2=N_{\max}/N_0}$ under the settings in Table 1. Thus, the actual type I error rate under the rule (6) is

$$\begin{aligned} &\alpha + E\{I(r>1)\Delta(r,t,z)\} \\ &\quad = \alpha + \int_{z_{CP=Q}}^{z_{\lambda^2=N_{\max}/N_0}} \Delta(r = N_{\max}/N_0, t, z)\varphi(z)dz + \int_{z_{\lambda^2=N_{\max}/N_0}}^{z_{\lambda=1}} \Delta(r=\lambda^2, t, z)\varphi(z)dz, \end{aligned}$$

where for any proposition A, $I(A)$ equals one if A is true and zero otherwise, and $\varphi(\cdot)$ is the probability density function for a standard normal variable. The above integrations were performed with Mathematica 6.0.

B. Calculation of the maximum type I error rate under the rule (7)

Let r_{peak} is the value of r, which gives the maximum value of $\Delta(r,t,z)$ given t and z in any r (>1). Since $\Delta(r,t,z)$ is an unimodal function of r given t and z, one can obtain $r_{\text{peak}} = t\left(\frac{z_\alpha}{z}\right)^2$ by solving $\frac{\partial}{\partial r}\Delta(r,t,z) = 0$. Note that r_{peak} is a decreasing function of z given t. When $r > r_{\min}$ under the rule (7), the value of $\Delta_{\max}(r,t,z)$ is determined according to the size of r_{peak} and $r_{\min}$, that is, r_{peak} gives the maximum if $r_{\min} \leq r_{\text{peak}}$, while $r_{\min}$ gives the maximum if $r_{\text{peak}} < r_{\min}$. Thus, the maximum type I error rate under the rule (7) is

$$\begin{aligned} \alpha_{\max} &= \alpha + E\{I(r > r_{\min})\Delta_{\max}(r,t,z)\} \\ &= \alpha + \int_{z_{CP=Q}}^{z_{r_{\text{peak}}=r_{\min}}} [I(z_{CP=Q} < z_{r_{\text{peak}}=r_{\min}})\Delta(r = r_{\text{peak}}, t, z) \\ &\qquad + \{1 - I(z_{CP=Q} < z_{r_{\text{peak}}=r_{\min}})\}\Delta(r = r_{\min}, t, z)]\varphi(z)dz \\ &\qquad + \int_{z_{r_{\text{peak}}=r_{\min}}}^{z_{\lambda=1}} \Delta(r = r_{\min}, t, z)\varphi(z)dz, \end{aligned}$$

where $z_{r_{\text{peak}}=r_{\min}} = z_\alpha\sqrt{t/r_{\min}}$ is the test statistic at $r_{\text{peak}} = t\left(\frac{z_\alpha}{z}\right)^2 = r_{\min}$. The above integrations were performed with Mathematica 6.0.

Japanese Journal of Biometrics Vol. 29, No. 2, 61–74 (2008)

Original Article

A Poisson Mixed Effects Model for Investigating the Exposure-by-Cohort Interaction: A Gibbs Sampling Approach

Seitaro Yoshida[*1], Yutaka Matsuyama[*1], Yasuo Ohashi[*1] and Hirotsugu Ueshima[*2]

[*1]Department of Biostatistics, School of Health Sciences and Nursing, University of Tokyo

[*2]Department of Health Science, Shiga University of Medical Science

e-mail:seitaro@epistat.m.u-tokyo.ac.jp

A meta-analysis is a useful method for taking the findings of many studies and combining them in the hopes of identifying consistent patterns and sources of disagreement among those findings. While we interpret the average exposure effect, it is necessary to examine the homogeneity of the observed exposure effects across cohort, that is, exposure-by-cohort interaction. If the homogeneity is confirmed, the conclusions concerning exposure effects can be generalized to a broader population. In this paper, a Poisson mixed effects model is used to investigate the cohort effects on the exposure as well as on the baseline risk. The marginal posterior distributions are estimated by a Markov Chain Monte Carlo method, i.e. the Gibbs sampling, to overcome current computational limitations. We illustrate the methods with analyses of data from the Japan Arteriosclerosis Longitudinal Study, in which the effects of smoking on stroke events are examined based on the individual data of 23,860 subjects among 10 cohorts.

Key words: exposure-by-cohort interaction, Generalizability, Gibbs sampling, Poisson mixed effects model.

1. Introduction

Smoking is known to be associated with an increased risk of cardiovascular disease (Peto 1994). Although many epidemiologic studies in Western populations have also identified smoking as an independent risk factor for stroke (Colditz et al. 1988; Wolf et al. 1988; Shinton and Beevers 1989), its relationship in Japanese people living in Japan remains inconclusive (Hirayama T 1981; Nakayama T et al. 1977; Kiyohara et al. 1990; Tanizaki et al. 2000; Yamagishi et al. 2003; Mannami et al. 2004; Ueshima et al. 2004; Iso et al. 2005).

A meta-analysis is a useful method for taking the findings of many studies and combining them, in the hopes of identifying consistent patterns and sources of disagreement among those findings. When conducting a meta-analysis, as many authors have stressed (Rothman et

Received July 2008. Revised August 2008. Accepted August 2008.

al. 2008), analysis of heterogeneity can be the most important function of meta-analysis. There is ordinarily no basis for assuming that the relative risk is constant across study cohorts. In fact, there are many situations that imply heterogeneity; for example, study cohorts are different not only in distributions of background factors such as age and sex, but also in environmental factors such as weather conditions and dietary habits, which are difficult to account for explicitly in the analysis.

The standard form of heterogeneity analysis is to regard the cohort as a stratification variable, that is, fixed effects, and to examine the exposure-by-cohort interaction by the analogy with ANOVA F-tests for interaction. However, small numbers of events per cohort and large numbers of confounders are often the case in most epidemiologic studies. In such cases, the loss of efficiency of analysis may be severe, if the cohort effects are taken to be fixed ones. Alternatively, in this article, we treat the cohort effects to be random ones in order to investigate the cohort effects on the exposure risk as well as on the baseline risk. The resulting statistical model for the observed data is a mixed effects model (Fitzmaurice et al. 2004; Greenland 2000), with the effects of exposure being fixed and the effects of cohort being random.

In some special cases such as the linear mixed effects model, the integral involved the likelihood function has a closed form, and ordinary iterative algorithms for maximizing the likelihood are used to obtain maximum likelihood or restricted maximum likelihood estimates for unknown model parameters (Laird and Ware 1982). For most nonlinear models such as the logistic and Poisson model, however, the likelihood function does not have an analytical form. Thus, one of the problems for fitting the generalized linear mixed effects model is the difficulty of the estimation of model parameters due to the requirement of the numerical integration techniques for calculation of the log-likelihood (Breslow and Clayton 1993). A variety of numerical approximation methods for maximizing the likelihood have recently been implemented in commercial software packages. For example, PROC NLMIXED in SAS directly maximizes an approximate integrated likelihood where the integration over the random effects is achieved using Gaussian quadrature (Pinheiro and Bates 1995). PROC GLIMMIX in SAS uses the linearization (Taylor series) methods known as restricted pseudo-likelihood estimation with an expansion around the current estimate of the best linear unbiased predictors of the random effects (Breslow and Clayton 1993; Wolfinger and O'Connell 1993).

As Fitzmaurice et al. (2004) have commented concerning the use of the above procedures, convergence of the algorithms should never be taken for granted, that is, neither should convergence to a global maximum be assumed. Their limited experience with these procedures indicates that it can be very sensitive to poor choices of starting values and/or the numerical accuracy of the quadrature used. In fact, in the analyses of the Japan Arteriosclerosis Longitudinal Study (JALS) data which are motivated example in this paper and are described in the next section, the algorithms by the NLMIXED/GLIMMIX procedures did not converge. For

this estimation problem with intractable high-dimensional integrals involved in the likelihood, Bayesian approaches have been proposed to avoid the need for numerical integration by taking repeated samples from the posterior distributions (Zeger and Karim 1991).

In this paper, we propose to use a Poisson mixed effects model to investigate the exposure-by-cohort interaction, and use a Gibbs sampling technique for model parameter inferences.

2. Methods

2.1 Study Population

The aim of the Japan Arteriosclerosis Longitudinal Study (JALS) is to investigate the risk factors for arteriosclerotic disease specific to Japanese population (Japan Arteriosclerosis Longitudinal Study (JALS) Group. 2008). The JALS is composed of two studies: one is a pooled study based on individual patient data of existing prospective cohort studies in Japan, which has been conducted between 1985 and December 2001, the other is a prospective cohort study, which is on-going since 2004. In this paper, we used data from the former study, in which a total of 66,691 men and women with baseline records on such as age (year), gender, BMI (body mass index) (kg/m^2) and history of smoking were registered in 17 regional cohorts and 4 occupational cohorts. The subjects had also several medical check-up data including medical history and laboratory tests.

In this paper, we excluded following subjects from the analyses; those who belonged to the occupational cohorts, because our focus was on the regional cohorts, those who belonged to the regional cohorts where data on stroke event was not observed, those aged under 40 or over 90 years old, because of the small numbers of subjects and stroke events in those age categories, those who with a history of stroke, and those who have missing data on a history of smoking and confounders. The final analysis population consisted of 9,087 men and 14,773 women in 10 cohorts. The endpoint was the first occurrence of stroke, comprising fatal and nonfatal event of ischemic stroke, hemorrhagic stroke and subarachnoid hemorrhage. All outcomes were classified according to the ICD-9 (9th revision of the International Classification of Diseases) until the end of 1994, and according to the ICD-10 since the beginning of 1995.

2.2 A Poisson Mixed Effects Model

We propose to use a Poisson mixed effects model to investigate a smoking-by-cohort interaction adjusted for age (< 65 or ≥ 65 years), SBP (systolic blood pressure; < 140 or ≥ 140 mmHg), BMI (< 25 or ≥ 25 kg/m^2), current alcohol drinking status (yes or no) and history of DM (diabetes mellitus; present or absent). All analyses were stratified by gender.

A form of the Poisson mixed effects model used for the analyses can be expressed as,

$$\log E(y_{ij} \mid b_i) = \log t_{ij} + \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + b_{0i} + b_{1i} x_{ij} \tag{1}$$

where y_{ij} represents the event variable ($y_{ij} = 1$ if stroke is observed, $y_{ij} = 0$ otherwise) for the jth subject ($j = 1, \ldots, n_i$) in the ith cohort ($i = 1, \ldots, N$); t_{ij} represents the person-year of follow-up;

x_{ij} represents a smoking status; z_{ij} represents a vector of covariates stated above; the parameters β and γ represent the fixed effects corresponding to x_{ij} and z_{ij}, respectively; b_{0i} and b_{1i} represent the random effects for the ith cohort, which are assumed to be normally distributed,

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}\right]. \tag{2}$$

In the model (1), the random effect b_{0i} is the deviation of the ith cohort from a baseline risk, while b_{1i} is the deviation of the ith cohort from an average smoking effect β_1, that is, smoking-by cohort interaction.

2.3 Full Conditional Distributions for Each Parameter

The Gibbs sampling is simply a technique for generating random variables from a difficult joint distribution indirectly without calculating the density. The mechanism is based only on elementary properties of Markov chains (Gelfand and Smith 1990). Gilks et al. (1993) have reviewed some applications of the Gibbs sampling to Bayesian models in medicine. To perform the Gibbs sampling, we require the full conditional distributions for each of the unknown parameters of the model.

In the Bayesian approach to analyzing the model (1), the parameters $\alpha = (\beta, \gamma)$ and D are random variables and treated symmetrically with the observed y and unobserved b_i. The marginal posterior distribution over the random effects corresponding to the model (1) is proportional to

$$f(\alpha, D \mid y) = \prod_{i=1}^{N} \int f(y_i \mid \alpha, b_i) \times |D|^{-1/2} \exp(-\frac{1}{2} b_i^T D^{-1} b_i) \pi(\alpha, D) db_i, \tag{3}$$

where $f(y_i \mid \alpha, b_i)$ is the conditional Poisson likelihood given b_i with mean expressed through (1), D is the variance-covariance matrix for b_i, and $\pi(\alpha, D)$ is the joint prior distribution for α and D. The necessary full conditional distributions for model (3) at iteration k ($k = 1, \ldots, K$) are given as follows.

2.3.1 Sampling of the Fixed Effects

Given the random effects $b_i^{(k)}$, the Poisson mixed effects model (1) reduces to the usual Poisson regression model with offset $b_{0i}^{(k)} + b_{1i}^{(k)} x_{ij}$ for each subject. If we assume a uniform prior for α, the full conditional posterior distribution for α is proportional to the Poisson likelihood function $\prod_{i=1}^{N} \prod_{j=1}^{n_i} f(y_{ij} \mid b_i^{(k)})$. In larger samples, this can be closely approximated by a multivariate normal distribution with mean vector $\hat{\alpha}^{(k)}$, the maximum likelihood estimate, and covariance matrix $\hat{V}_{\hat{\alpha}}^{(k)}$, the inverse of the corresponding observed Fisher information. That is, to sample from full conditional distributions for α at iteration k, we find $\hat{\alpha}^{(k)}$ and $\hat{V}_{\hat{\alpha}}^{(k)}$ by performing Poisson regression of y_{ij} on x_{ij} using the simulated (fixed) values $b_{0i}^{(k)} + b_{1i}^{(k)} x_{ij}$ as offsets and generate a random variate $\hat{\alpha}^{(k+1)}$ from a multivariate normal distribution $N(\hat{\alpha}^{(k)}, \hat{V}_{\hat{\alpha}}^{(k)})$.

2.3.2 Sampling of the Random Effects

The full conditional posterior distribution for b_i dose not have a closed form and we cannot draw samples directly from its distribution, because it involves the same intractable integral with respect to b_i that we avoided in the maximum likelihood analysis. The distribution can be expressed as

$$\frac{f(y_i \mid \alpha, b_i)g(b_i \mid D)\pi(\alpha, D)}{\int f(y_i \mid \alpha, b_i)g(b_i \mid D)\pi(\alpha, D)db_i}, \tag{4}$$

where $g(b_i \mid D)$ is the normal density given in (2). Although the numerator in (4) is easily evaluated, the scale factor in the denominator cannot be evaluated explicitly. In Gibbs sampling, however, only a random variable from the conditional distribution is needed. It can be obtained using random variate generating methods such as rejection sampling without evaluating the integral. We used the normal rejection sampling method of Zeger and Karim (1991). Their method is to find the mode and curvature of the numerator of (4), call it $p_i(b_i)$, and to match a Gaussian kernel with the covariance matrix multiplied by an inflation factor to $p_i(b_i)$. This Gaussian kernel is then rescaled to be at least as large as $p_i(b_i)$ at the mode, and used as the envelope function. We used SAS/IML nonlinear optimization subroutines, Newton-Raphson optimization, to find the mode and curvature of $p_i(b_i)$, and a variance inflation factor of 2.5 was used.

2.3.3 Sampling of the Random Effects Variance

We have assumed that the random effects b_i are independent normal $N(0, D)$ random variables. A non-informative prior for D is $P(D) \propto |D|^{-\frac{q+1}{2}}$, where $q = 2$ in the model (1). Then the full conditional posterior distribution of D^{-1} follows a Wishart distribution with parameter matrix $(\sum_{i=1}^{N} b_i b_i^T)^{-1}$ with $(N - q + 1)$ df (Gelman et al. 2004). Simulation from the Wishart distribution for 2×2 matrix D^{-1} is easily accomplished using the algorithm of Odell and Feiveson (1966). Their algorithm is as follows: with $Ga(\cdot, \cdot)$ and $N(\cdot, \cdot)$ denoting gamma and normal distributions, respectively, draw independently from $U_1 \sim Ga(\frac{\nu}{2}, \frac{1}{2})$, $U_2 \sim Ga(\frac{\nu-1}{2}, \frac{1}{2})$, and $U_3 \sim N(0, 1)$, set

$$W = \begin{pmatrix} U_1 & U_3\sqrt{U_1} \\ U_3\sqrt{U_1} & U_2 + U_3^2 \end{pmatrix},$$

then if $S^{-1} = (H^{1/2})^T(H^{1/2})$, $D^{-1} = (H^{1/2})^T W (H^{1/2}) \sim Wishart(S^{-1}, \nu)$.

2.4 Implementation of the Gibbs Sampling

The estimation of model parameters was conducted using SAS/IML (SAS Institute Inc., Cary NC). We simulated independent sequences of length 2,000 with three kinds of starting values, where the procedure entailed sampling from 1,000 post-convergence iterations in each sequence, thus yielding 3,000 total iterations upon which we based our posterior estimates. We set three kinds of initial values of the random effects variances (d_{11}, d_{12}, d_{22}) at $(1.0, 0, 0.1)$,

(0.1, 0, 0.6), (0.1, 0, 1.0), respectively. First one represents the large baseline risk and small interaction effect, second one represents small baseline risk and medium interaction effect, and third one represents small baseline risk and large interaction effect. For the random effects, initial values were drawn from the multivariate normal distribution (2) with each initial value of the random effects variances.

Convergence of the Markov chains was assessed by the Gelman-Rubin statistic, that is, potential scale reduction (PSR) factor (Gelman and Rubin 1993). Details of the PSR are given in Appendix.

3. Results

Table 1 and 2 show the baseline and stroke events characteristics of the analysis 10 cohorts, respectively. The total number of stroke events was 345 among 9,087 men (3.8%) and 368 among 14,773 women (2.5%). The total incidence rate (/year) for stroke events was 4.59×10^{-3} in men and 2.90×10^{-3} in women, where substantial variations across cohorts were seen in women. The incidence rate ratios (IRRs) for current smoking were not consistent across cohorts, particularly, in women.

Table 3 shows the posterior summaries of the model parameters in (1). The last column in Table 3 shows the estimate of PSR comparing three runs for each of several model parameters. The estimates were all near 1, indicating that convergence occurred for these parameters. Table 4 shows the estimates of IRR for each fixed effects parameter. After adjustment of age, SBP, BMI, current alcohol drinking status and history of DM, an association between current smoking status and stroke events was observed among men (IRR = 1.55, 95% confidence region (CR) = 1.22-2.02), while the significant association was not observed among women (IRR = 1.32, 95% CR = 0.76-2.18).

In Table 4, results from the ordinary Poisson model are also shown, that is, estimates of IRR based on the fixed effects model where random effect terms were excluded from the model (1). The posterior features for smoking effects changed little compared with the ordinary model ignoring cohort effects. This indicates that incorporating cohort effects was not critical for drawing conclusions on the overall smoking effects. However, the mixed effects model can be used to evaluate the cohort effects on the smoking risk as well as on the baseline risk.

The posterior summaries of the random effects variances, which represent the volumes of cohort differences, are also shown in Table 3. For both genders, baseline variability (d_{11}) was larger than variability of smoking effect (d_{22}). The variances in women were larger than those of men in both risks. Because the random effects were assumed to be normally distributed in (2), the posterior mean of the variance of random effects indicates that 95% of cohorts have the baseline risk ($\exp(\pm 1.96\sqrt{\hat{d}_{11}})$) in the range 0.48 to 2.08 for men and 0.34 to 2.98 for women. Likewise, the variability of smoking-by-cohort interaction ($\exp(\pm 1.96\sqrt{\hat{d}_{22}})$) was in the range

Table 1. Characteristics of analysis 10 cohorts stratified by gender.

Gender	Cohort	Number of subjects	Follow-up period(year)	Age(year)		SBP(mmHg)		BMI(kg/m^2)		Current smoking		Current drinking		DM	
		N	Mean	Mean	SD	Mean	SD	Mean	SD	N	%	N	%	N	%
Men	1	670	6.2	59.9	10.6	132	19.7	23.7	3.2	341	50.9	448	66.9	40	6.0
	2	466	11.3	56.3	13.3	134	16.3	23.5	2.9	229	49.1	289	62.0	7	1.5
	3	1086	10.4	55.8	10.8	137	20.1	22.9	2.9	679	62.5	839	77.3	113	10.4
	4	635	9.9	58.7	11.7	134	16.3	23.5	3.0	319	50.2	410	64.6	119	18.7
	5	1840	9.8	54.3	9.3	137	19.7	23.3	3.0	1150	62.5	1254	68.2	207	11.3
	6	1261	9.2	55.9	11.1	134	20.1	22.6	2.8	733	58.1	874	69.3	108	8.6
	7	1182	7.1	56.3	13.6	133	19.3	22.3	3.0	688	58.2	860	72.8	100	8.5
	8	977	5.0	58.3	12.3	132	17.7	23.2	2.9	391	40.0	652	66.7	77	7.9
	9	292	9.4	60.8	10.3	134	20.4	22.7	3.1	138	47.3	173	59.2	16	5.5
	10	678	3.7	63.1	10.6	136	20.3	23.2	3.0	250	36.9	499	73.6	47	6.9
	Total	9087	8.3	57.1	11.6	135	19.3	23.0	3.0	4918	54.1	6298	69.3	834	9.2
Women	1	909	6.4	59.4	9.8	132	21.3	23.8	3.2	66	7.3	153	16.8	33	3.6
	2	958	12.1	55.7	11.9	131	15.2	24.3	3.5	15	1.6	13	1.4	6	0.6
	3	1439	10.8	56.0	10.7	133	18.9	24.0	3.4	31	2.2	41	2.8	77	5.4
	4	1204	10.7	57.2	10.4	130	16.4	24.6	3.4	29	2.4	90	7.5	187	15.5
	5	2510	10.1	54.7	9.5	136	19.7	23.9	3.4	165	6.6	201	8.0	152	6.1
	6	2438	9.8	52.7	10.6	131	20	22.6	3.0	259	10.6	372	15.3	99	4.1
	7	1661	7.3	56.4	14.1	131	19.9	22.8	3.4	151	9.1	402	24.2	66	4.0
	8	2200	5.3	52.3	13.6	124	19	22.8	3.2	93	4.2	318	14.5	66	3.0
	9	455	9.8	60.6	10.1	132	22.3	23.0	3.1	19	4.2	42	9.2	18	4.0
	10	999	3.8	61.9	10.9	132	20.9	23.0	3.2	20	2.0	259	25.9	38	3.8
	Total	14773	8.6	55.6	11.7	131	19.6	23.4	3.3	848	5.7	1891	12.8	742	5.0

SBP: Systolic blood pressure; BMI: Body mass index; DM: Diabetes mellitus

Table 2. Stroke events of analysis 10 cohorts stratified by gender.

Gender	Cohort	Stroke events N	Stroke events %	Person-time (year)	Incidence rate (/year)	IRR for Current smoking
Men	1	27	4.0	4173	6.47×10^{-3}	0.79
	2	22	4.7	5278	4.17×10^{-3}	1.29
	3	56	5.2	11284	4.96×10^{-3}	1.44
	4	59	9.3	6285	9.39×10^{-3}	1.40
	5	68	3.7	17953	3.79×10^{-3}	1.83
	6	38	3.0	11633	3.27×10^{-3}	1.08
	7	31	2.6	8413	3.68×10^{-3}	2.03
	8	18	1.8	4925	3.65×10^{-3}	0.94
	9	14	4.8	2747	5.10×10^{-3}	1.19
	10	12	1.8	2542	4.72×10^{-3}	1.23
	total	345	3.8	75233	4.59×10^{-3}	1.29
Women	1	27	3.0	5792	4.66×10^{-3}	0.52
	2	24	2.5	11553	2.08×10^{-3}	6.39
	3	58	4.0	15520	3.74×10^{-3}	0.90
	4	66	5.5	12940	5.10×10^{-3}	1.26
	5	70	2.8	25261	2.77×10^{-3}	1.35
	6	34	1.4	23915	1.42×10^{-3}	2.36
	7	38	2.3	12166	3.12×10^{-3}	0.91
	8	16	0.7	11664	1.37×10^{-3}	0.00
	9	31	6.8	4457	6.96×10^{-3}	0.77
	10	4	0.4	3782	1.06×10^{-3}	0.00
	total	368	2.5	127050	2.90×10^{-3}	1.15

IRR: Incidence rate ratio

Table 3. Posterior summaries for the model parameters in (1).

Gender	Parameter	Mean	Standard deviation	2.5 percentile	97.5 percentile	PSR
Men	Intercept	-6.39	0.24	-6.85	-5.91	1.00
	Current smoking (yes vs no)	0.44	0.13	0.20	0.70	1.00
	Age ($\geq$ 65 years vs $<$ 65)	1.03	0.11	0.81	1.25	1.00
	SBP ($\geq$ 140 mmHg vs $<$ 140)	0.88	0.11	0.65	1.10	1.00
	BMI ($\geq$ 25 kg/m^2 vs $<$ 25)	0.03	0.13	-0.22	0.28	1.00
	Current drinking (yes vs no)	-0.04	0.08	-0.21	0.13	1.00
	DM (present vs absent)	0.57	0.14	0.29	0.85	1.00
	d_{11}	0.14	0.23	0.01	0.63	1.01
	d_{12}	-0.01	0.07	-0.11	0.05	1.01
	d_{22}	0.03	0.10	0.00	0.17	1.14
Women	Intercept	-6.72	0.22	-7.15	-6.28	1.04
	Current smoking (yes vs no)	0.28	0.27	-0.28	0.78	1.04
	Age ($\geq$ 65 years vs $<$ 65)	1.20	0.12	0.97	1.43	1.04
	SBP ($\geq$ 140 mmHg vs $<$ 140)	0.87	0.10	0.67	1.07	1.04
	BMI ($\geq$ 25 kg/m^2 vs $<$ 25)	0.22	0.11	0.00	0.44	1.04
	Current drinking (yes vs no)	-0.08	0.19	-0.45	0.31	1.04
	DM (present vs absent)	0.24	0.17	-0.11	0.58	1.04
	d_{11}	0.31	0.35	0.03	1.19	1.01
	d_{12}	-0.02	0.19	-0.48	0.30	1.00
	d_{22}	0.18	0.39	0.00	1.07	1.04

SBP: Sytolic blood pressure; BMI: Body mass index; DM; Diabetes mellitus
PSR: Potential scale reduction

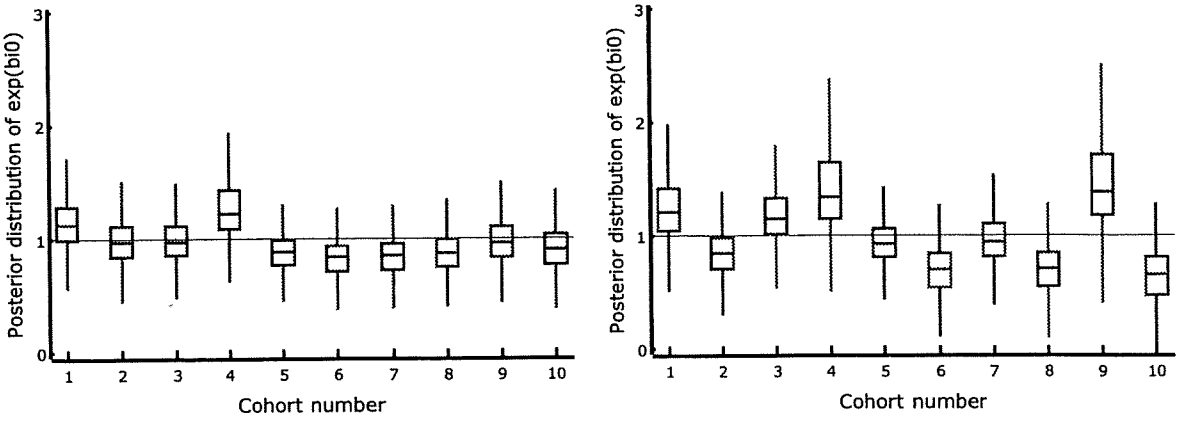
Table 4. Comparison of the results between mixed effects and ordinary Poisson model.

Gender	Parameter	Poisson mixed effexcts model		Ordinary Poisson model	
		IRR	95% CR	IRR	95% CR
Men	Current smoking (yes vs no)	1.55	1.22-2.02	1.53	1.22-1.91
	Age ($\geq$ 65 years vs $<$ 65)	2.79	2.24-3.49	2.86	2.29-3.57
	SBP ($\geq$ 140 mmHg vs $<$ 140)	2.41	1.91-3.01	2.38	1.91-2.96
	BMI ($\geq$ 25 kg/m^2 vs $<$ 25)	1.03	0.80-1.32	1.05	0.82-1.35
	Current drinking (yes vs no)	0.96	0.81-1.14	0.96	0.76-1.20
	DM (present vs absent)	1.76	1.34-2.33	1.80	1.34-2.40
Women	Current smoking (yes vs no)	1.32	0.76-2.18	1.30	0.85-2.00
	Age ($\geq$ 65 years vs $<$ 65)	3.32	2.64-4.17	3.43	2.77-4.24
	SBP ($\geq$ 140 mmHg vs $<$ 140)	2.39	1.96-2.93	2.35	1.90-2.91
	BMI ($\geq$ 25 kg/m^2 vs $<$ 25)	1.25	1.00-1.56	1.32	1.07-1.63
	Current drinking (yes vs no)	0.92	0.64-1.36	0.88	0.60-1.29
	DM (present vs absent)	1.27	0.90-1.78	1.37	0.97-1.95

IRR: Incidence rate ratio; CR: Confidence region; CI: Confidence interval
SBP: Sytolic blood pressure; BMI: Body mass index; DM; Diabetes mellitus

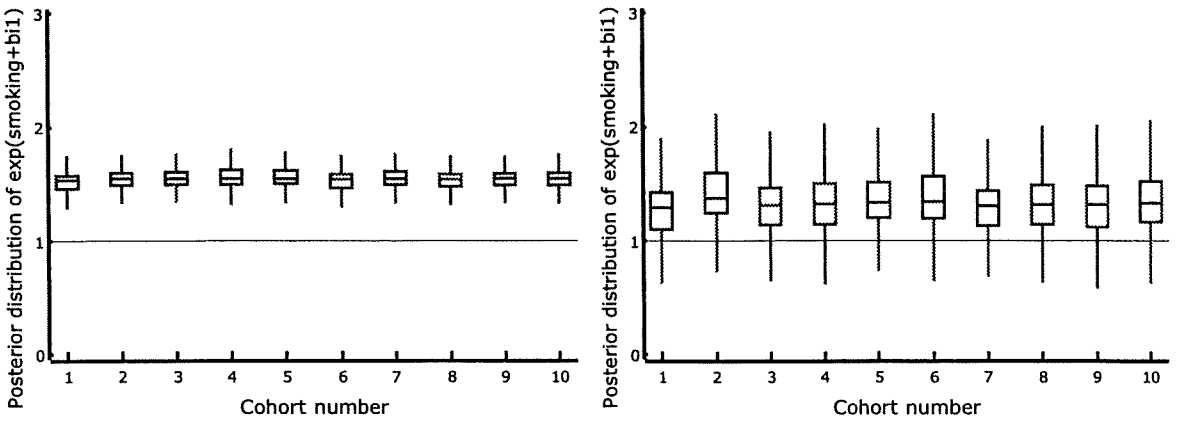
0.71 to 1.40 and 0.44 to 2.30, respectively.

Figures 1 and 2 summarize the posterior distributions of the cohort effects for the baseline risk and smoking effects, respectively. These are the box plots of the posterior sample values of generated ones for each of the 10 cohorts. The baseline risk, $\exp(b_{0i})$, is the deviation in the ith cohort from an overall baseline risk, and the smoking effect, $\exp(\beta_1 + b_{1i})$, is the deviation in the ith cohort from an overall smoking effect. Figure 1 indicates substantial variation in the baseline risk across cohorts, particularly in women. The baseline risk seems to vary considerably across cohorts (prior mean of the baseline risk is one). On the other hand, Figure 2 appears to indicate the homogeneity in the effect of smoking across cohorts in both genders, that is, there is little smoking-by-cohort interaction in this data. Thus, there is little difference in the smoking effects



(a) Men (b) Women

Fig. 1. Posterior distribution of baseline risk ($\exp(b_{0i})$) in each cohort.



(a) Men (b) Women

Fig. 2. Posterior distribution of smoking effect ($\exp(\beta_1 + b_{1i})$) in each cohort (smoking-by cohort interaction).

on stroke events across cohorts and the smoking is shown to be an independent risk factor for the events, while there appears to be substantial variation in the baseline risk across cohorts. This result indicates that the observed smoking effects might be generalized to a broader population.

4. Discussion

In this paper, we proposed to use a Poisson mixed effects model (1) with two random effects to investigate the exposure-by-cohort interaction. It is important to investigate the cohort effects on the exposure risk in addition to the baseline risk. This model is useful not only for the meta-analysis of individual epidemiologic data like in the JALS, but also for the analysis of multicenter clinical trials (Matsuyama et al. 1998).

Until recently, a potential limitation of generalized linear mixed models was their computational burden. Because there is no simple closed-form solution for the marginal likelihood, numerical integration (Pinheiro and Bates 1995) or pseudo-likelihood (Breslow and Clayton 1993; Wolfinger and O'Connell 1993) techniques are required. Maximum or restricted maximum likelihood estimation has only recently been implemented in standard statistical software, for example, PROC NLMIXED or PROC GLIMMIX in SAS. However, problems with convergence are likely to arise when complicated models with two or three random effects are fitted (Fitzmaurice et al. 2004; Evans et al. 2001). This non-convergence problem seems to frequently occur in highly unbalanced data or in sparse data, that is, the number of events is small relative to the adjustment variables. In fact, in our data analysis of the JALS data, the number of stroke events was small in each cohort as shown in Table 2, and the optimization algorithms by the NLMIXED/GLIMMIX procedures for the model (1) did not converge, although a number of things was tried, for example, change the initial values by using a grid search specification to obtain a set of good feasible starting values, change or modify the update or optimization