A method of assessing the sensitivity of the Cochran–Mantel–Haenszel test to an unobserved confounder

BY BINBING YU1,* AND JOSEPH L. GASTWIRTH2

1Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, Bethesda, MD 20892, USA
2Department of Statistics, The George Washington University, Washington, DC 20037, USA

Observational studies, including the case-control design frequently used in epidemiology, are subject to a number of biases and possible confounding factors. Failure to adjust with them may lead to an erroneous conclusion about the existence of a causal relationship between exposure and disease. The Cochran–Mantel–Haenszel (CMH) test is widely used to measure the strength of the association between an exposure and disease or response, after stratifying on the observed covariates. Thus, observed confounders are accounted for in the analysis. In practice, there may be causal variables that are unknown or difficult to obtain. Hence, they are not incorporated into the analysis. Sensitivity analysis enables investigators to assess the robustness of the findings. A method for assessing the sensitivity of the CMH test to an omitted confounder is presented here. The technique is illustrated by re-examining two datasets: one concerns the effect of maternal hypertension as a risk factor for low birth weight infants and the other focuses on the risk of allopurinol on having a rash. The computer code performing the sensitivity analysis is provided in appendix A.

Keywords: Cochran–Mantel–Haenszel test; confounders; sensitivity analysis; stratification

1. Introduction

Observational studies are commonly used to evaluate the effect of treatments or chemical exposure, e.g. the association between risk factors and a subsequent event. Unlike randomized experiments, researchers cannot control the assignment of treatments to subjects to ensure that subjects receiving different treatments or exposure are similar with respect to major covariates. The Cochran–Mantel–Haenszel (CMH; Cochran 1954; Mantel & Haenszel 1959) test is widely used to test the strength of the association after controlling for the observed confounders. The data are stratified into multiple 2×2 tables by known confounders, where in each stratum the distributions of the confounders in both groups are similar.

* Author for correspondence (yubi@mail.nih.gov).

One contribution of 13 to a Theme Issue ‘Mathematical and statistical methods for diagnoses and therapies’.
Since observational studies remain subject to unobserved confounding, one
should explore how sensitive the results are to a possible omitted variable.
Extensive work has been done on how to perform sensitivity analysis for
observational data by Cornfield et al. (1959), Rosenbaum & Rubin (1983, 1984),
Rosenbaum & Krieger (1990), Gastwirth (1992), Rosenbaum (2002) and others.
Most of the literature models the association between an unobserved confounder,
denoted by $U$, and treatment assignment, and then recomputes the test statistics
and $p$ values for a range of plausible differences in the prevalences of the
confounder in the two groups. If a moderate degree of confounding could change
the inference about the treatment effect on the outcome, then one questions the
soundness of the conclusions. Gastwirth et al. (1998) proposed a dual and
simultaneous sensitivity analysis for matched pair data, which extended a
strengthened form of the original Cornfield inequality (Gastwirth 1988, p. 296).
Both the relationship between $U$ and treatment assignment and that between $U$ and
response are jointly modelled. This simultaneous form allows one to incorporate
subject matter knowledge about both relationships into the sensitivity analysis.

This paper extends the ‘simultaneous’ approach to examine the sensitivity of
the CMH test to an unobserved confounder. The rest of this paper is organized as
follows. Section 2 reviews the CMH test and describes the stratified model for the
simultaneous relationships. Then we apply the sensitivity analysis to a dataset
examining maternal hypertension as a risk factor for low birth weight (LBW)
infants and another dataset concerned with a side effect (rash) of allopurinol in
§3. We discuss the potential use and future development of sensitivity analysis in
observational studies in §4.

2. Sensitivity analysis for the CMH test

Let $X$ be the exposure, e.g. environmental hazard, genetic factor, radiation
exposure; and $Y$ be the outcome, e.g. disease, mortality, or in an employment
discrimination case, being promoted or passing a pre-employment exam. We are
interested in testing the causal relationship between $X$ and $Y$ using the CMH test
to control for other known factors related to the response. Usually the data are
organized into $J$ $2 \times 2$ tables, where in each stratum the covariates have similar
values. The data for the $j$th stratum ($j=1, \ldots, J$) are shown below.

| exposure group $X$    | outcome $Y$                      |       |
|-----------------------|----------------------------------|-------|
|                       | event (1)                        | non-event (0) | total |
| exposed (1)           | $a_j$                            | $b_j$  | $m_{1j}$ |
| not exposed (0)       | $c_j$                            | $d_j$  | $m_{0j}$ |
| total                 | $n_{1j}$                         | $n_{0j}$ | $N_j$  |

Let $\pi_{1j}(\pi_{0j}), j=1, \ldots, J$, be the probability that subjects who are exposed (not
exposed) in stratum $j$ has an event. In terms of these parameters, the odds ratio
between exposure and outcome in stratum $j$ is

$$
\theta_j = \pi_{1j}(1 - \pi_{0j})/[(1 - \pi_{0j})(1 - \pi_{1j})], \quad j = 1, \ldots, J.
$$

*Phil. Trans. R. Soc. A* (2008)
Furthermore, assume that the odds ratio has a common value $\theta$. Let $D = \{a_j, b_j, c_j, d_j, j = 1, \ldots, J\}$ be the observed data for the $J \times 2 \times 2$ tables. When the responses for each group in all strata follow independent binomial distributions, the full data have the product binomial likelihood function

$$L(\theta|D) = \prod_{j=1}^{J} \left( \frac{m_{1j}}{a_j} \right) \pi_{1j}^{a_j}(1 - \pi_{1j})^{b_j} \left( \frac{m_{0j}}{c_j} \right) \pi_{0j}^{c_j}(1 - \pi_{0j})^{d_j}.$$  \hspace{1cm} (2.1)

From (2.1), we have $E(a_j) = m_{1j}\pi_{1j}$, $V(a_j) = m_{1j}\pi_{1j}(1 - \pi_{1j})$, $E(c_j) = m_{0j}\pi_{0j}$ and $V(c_j) = m_{0j}\pi_{0j}(1 - \pi_{0j})$.

If all the confounders are observed and controlled by stratification, then the null hypothesis of no-exposure effect is $H_0: \pi_{0j} = \pi_{1j}$, $j = 1, \ldots, J$ or $H_0: \theta = 1$. Cochran’s (1954) statistic for testing $H_0: \theta = 1$ versus $H_1: \theta > 1$ is

$$W_0 = \sqrt{\frac{\sum_{j=1}^{J} w_j(p_{1j} - p_{0j})}{\sum_{j=1}^{J} w_j(1 - \tilde{p}_j)}}.$$  \hspace{1cm} (2.2)

where $w_j = m_{1j}m_{0j}/N_j$, $\tilde{p}_j = (m_{1j}p_{1j} + m_{0j}\tilde{p}_{0j})/N_j$ and $p_{1j} = a_j/m_{1j}$, $p_{0j} = c_j/m_{0j}$ are the estimates of $\pi_{1j}$ and $\pi_{0j}$, respectively. Using the conditional central hypergeometric likelihood rather than the product binomial (2.1), Mantel & Haenszel (1959) proposed a similar test. The two tests are asymptotically equivalent and are often referred to as the CMH test. The CMH test is known to be the optimal method of combining multiple $2 \times 2$ tables with common odds ratio (Woolson et al. 1986).

The large sample distribution of (2.2) converges to a standard normal distribution under the null hypothesis that $\theta = 1$ (Cochran 1954). More generally, the statistic

$$W = \sqrt{\frac{\sum_{j=1}^{J} w_j[(p_{1j} - p_{0j}) - (\pi_{1j} - \pi_{0j})]}{\sum_{j=1}^{J} w_j^2 \left[ \frac{\pi_{1j}(1 - \pi_{1j})}{m_{1j}} + \frac{\pi_{0j}(1 - \pi_{0j})}{m_{0j}} \right]}}.$$  \hspace{1cm} (2.3)

converges to a standard normal distribution as $N_j \to \infty$ (Woolson et al. 1986).

Sometimes, information on known risk factors is not available or other risk factors related to the response have not been discovered. Here we assume that there is an unobserved factor $U$ and that the relationship between binary outcome $Y$ and exposure $X$ and confounder $U$ is modelled by

$$\pi_{xj|u} = P(Y = 1|X = x, S = j, U = u) = \frac{\exp(\alpha_j + \beta x + \gamma u)}{1 + \exp(\alpha_j + \beta x + \gamma u)},$$ \hspace{1cm} (2.4)

where $\gamma$ is the effect of $U$, which is called the strength parameter (Yu & Gastwirth 2005), and $\beta$ is the true effect of exposure $X$. The null hypothesis of no-exposure effect is $H_0: \beta = 0$ and the alternative hypothesis is $H_1: \beta \neq 0$. 

**Phil. Trans. R. Soc. A** (2008)
The unobserved variable $U$ may also be associated with the exposure. Assuming that $U$ is binary, the prevalence of $U$ by exposure level in each stratum is modelled by

$$P(U = 1 | X = x, S = j) = \frac{\exp(\lambda_j + \delta_j x)}{1 + \exp(\lambda_j + \delta_j x)}.$$ \hspace{1cm} (2.5)

Here, the $\delta_j$, indicating the association between $X$ and $U$ in each stratum, are called the imbalance parameters (Yu & Gastwirth 2005). To simplify the calculation, we assume that the imbalance parameters $\delta_j$ are equal to $\delta$. Then, $q_j = e^\delta/(1 + e^\delta)$ is the prevalence of $U$ in the non-exposed group in stratum $j$.

When the confounder $U$ is omitted in the analysis, the marginal probability of response at each exposure level $x$ in stratum $j$ becomes

$$\pi_{xj} = \frac{\exp(\alpha_j + \beta x + \gamma) \exp(\lambda_j + \delta x)}{1 + \exp(\alpha_j + \beta x + \gamma) + \exp(\lambda_j + \delta x)} + \frac{\exp(\alpha_j + \beta x)}{1 + \exp(\alpha_j + \beta x) + \exp(\lambda_j + \delta x)}.$$ 

The true null hypothesis of no-exposure effect is $H_0$: $\beta = 0$, which is really the hypothesis of interest. Note that $H_0$: $\pi_{0j} = \pi_{1j}$, $j = 1, ..., J$, and $H_0$: $\beta = 0$ are not equivalent when there is an omitted confounder. If $U$ is not a confounder, i.e. $\delta = 0$ or $\gamma = 0$, then the equalities $\pi_{1j} = \pi_{0j}$, $j = 1, ..., J$, still hold under the null hypothesis $H_0$: $\beta = 0$, so that the CMH test $C_0$ is indeed testing the effect of exposure. When $U$ is indeed a confounder, it has been shown that, if both $\gamma > 0$ and $\delta > 0$, then $\pi_{1j} = \pi_{0j}$, $j = 1, ..., J$, even when the true null hypothesis holds (Yu & Gastwirth 2005) and the CMH test would be biased. This implies that confounder $U$ could induce spurious positive association between exposure and outcome. However, if $\gamma > 0$ and $\delta < 0$ then the true association between $X$ and $Y$ will be underestimated if the confounder is omitted.

If the strength parameter $\gamma$, the imbalance parameter $\delta$ and the prevalences $q_1, ..., q_J$ of the confounder are known or can be reliably estimated from other sources, the estimates of $\pi_{1j}$ and $\pi_{0j}$ can be calculated from the likelihood (2.1) under the null hypothesis $H_0$: $\beta = 0$ (see appendix A). Then the correct CMH statistic is obtained by substituting the estimates $(\hat{\pi}_{1j}, \hat{\pi}_{0j})$ for $(\pi_{0j}, \pi_{1j})$ in (2.3) as the effect of $U$ is now incorporated into the response probabilities $\pi_{xj}$. In practice these quantities are unknown, so one assumes a possible range for $(q_1, ..., q_J, \gamma, \delta)$; the test statistic and the $p$ value are calculated using (2.3). A table showing the $p$ values with respect to different parameters for confounding can be used to assess the plausibility that the current inference could be altered due to the effect of an unobserved factor.

3. Application

The proposed sensitivity analysis will be applied to data from two observational studies. In these applications, $C = \exp(\gamma)$ measures the effect of $U$ on binary response $Y$ and $D = \exp(\delta)$ measures the increasing odds of $U$ being positive when $X$ increases by one unit.
Sensitivity analysis of maternal hypertension as a risk factor for LBW infants

Data on 500 singleton births in a London Hospital (Hills & De Stavola 2002) were used to examine the sensitivity of the association of maternal hypertension as a risk factor for LBW infants. The dataset provides the birth information of 500 singletons with the following eight variables: identity number for mother and baby, birth weight of baby (indicator for birth weight less than 2500 g), gestation period (indicator for gestation period less than 37 weeks), maternal age, indicator for maternal hypertension and sex of baby (1, male; and 2, female). Ten mothers with missing gestation period information were dropped from the analysis. The exposure of interest (X) is maternal hypertension and the event of interest is LBW infant (Y).

‘Advanced’ maternal age is defined as any expectant mother who will have reached her 35th birthday by delivery date. The mothers are divided into ‘young’ and ‘advanced’ age group according to this criterion. The number of infants with normal birth weight (NBW) and LBW by sex, gestation status, maternal age and maternal hypertension are shown in Table 1. The CMH test is significant with \( p = 0.015 \) and the odds ratio is 2.74 with 95% CI (1.23, 6.14). The Breslow–Day test for homogeneity shows that the odds ratio estimates are homogeneous across different strata (\( p = 0.785 \)). Is it safe to draw the conclusion that maternal hypertension is causally related to LBW?

Despite substantial reductions in US infant mortality during the past several decades, black–white disparities in infant mortality rates persist. Important determinants of racial/ethnic differences in infant mortality are LBW, defined as less than 2500 g (US Department of Health and Human Services 2000). The number of LBW infants among 1000 live births for the whites are 5.7–6.5 and for the blacks are 12.7–13.0 from 1980 to 2000 (Centers for Disease Control and Prevention 2002; http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5127a1.htm). Hence, race is an important risk factor with an apparent odds ratio of approximately 2.0–2.4 for LBW (David & Collins 1997). A recent study (Geronimus et al. 2007) also showed that being black increases the odds ratio of hypertension by 2.11–4.04 for women between 15 and 65 years of age. Also, LBW is related to many other risk factors, e.g. the use of assisted reproductive technology, use of alcohol or drugs. Here, we focus on the race as the unobserved confounder (U).

Table 1. Infants’ birth weight by sex, gestation status, maternal age and hypertension.

| sex         | gestation | maternal age | normal blood pressure | hypertension |
|-------------|-----------|--------------|-----------------------|--------------|
|             |           |              | NBW | LBW | NBW | LBW |
| male        | termed    | young        | 98  | 2   | 15  | 2   |
|             |           | advanced     | 87  | 6   | 13  | 2   |
|             | preterm   | young        | 7   | 5   | 2   | 4   |
|             |           | advanced     | 7   | 2   | 1   | 3   |
| female      | termed    | young        | 91  | 5   | 13  | 1   |
|             |           | advanced     | 83  | 3   | 6   | 0   |
|             | preterm   | young        | 4   | 6   | 2   | 5   |
|             |           | advanced     | 2   | 11  | 0   | 2   |

(a) Sensitivity analysis of maternal hypertension as a risk factor for LBW infants

Despite substantial reductions in US infant mortality during the past several decades, black–white disparities in infant mortality rates persist. Important determinants of racial/ethnic differences in infant mortality are LBW, defined as less than 2500 g (US Department of Health and Human Services 2000). The number of LBW infants among 1000 live births for the whites are 5.7–6.5 and for the blacks are 12.7–13.0 from 1980 to 2000 (Centers for Disease Control and Prevention 2002; http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5127a1.htm). Hence, race is an important risk factor with an apparent odds ratio of approximately 2.0–2.4 for LBW (David & Collins 1997). A recent study (Geronimus et al. 2007) also showed that being black increases the odds ratio of hypertension by 2.11–4.04 for women between 15 and 65 years of age. Also, LBW is related to many other risk factors, e.g. the use of assisted reproductive technology, use of alcohol or drugs. Here, we focus on the race as the unobserved confounder (U).
because it is a recognized risk factor. We assume that the mother’s race is not related to the sex of their babies, but that the percentage of mothers who are black varies across the maternal age groups as well as gestation term. The possible scenarios for the percentage of black mothers by maternal age and gestation term are shown in Table 2. In scenarios 1 and 2, we assume that mothers with preterm babies consist of a higher percentage of black women. In scenarios 3 and 4, we assume that there are more black mothers with normal maternal age. In scenario 5, the black mothers tend to be younger and have more preterm babies than whites.

Table 2. The prevalences of black mothers by maternal age and gestation term for five possible scenarios.

| gestation | maternal age | prevalence | scenario (%) |
|-----------|--------------|------------|--------------|
|           |              |            | 1  | 2  | 3  | 4  | 5  |
| preterm   | normal       | $q_1$      | 10 | 5  | 10 | 20 | 10 |
|           | advanced     | $q_2$      | 10 | 5  | 5  | 10 | 5  |
| preterm   | normal       | $q_3$      | 20 | 10 | 10 | 20 | 20 |
|           | advanced     | $q_4$      | 20 | 10 | 5  | 10 | 10 |

Table 3. The effect of $U$ on the $p$ value of the CMH test for different values of $C$ and $D$.

| scenario | $C=\exp(\gamma)$ | $D=\exp(\delta)$ |
|----------|-------------------|-------------------|
|          | 2  | 3  | 4  | 5  | 2  | 3  | 4  | 5  | 2  | 3  | 4  | 5  | 2  | 3  | 4  | 5  | 2  | 3  | 4  | 5  |
| 1        | 2  | 0.064 | 0.079 | 0.094 | 0.107 | 0.057 | 0.066 | 0.074 | 0.083 | 0.065 | 0.083 | 0.101 | 0.119 | 0.072 | 0.099 | 0.127 | 0.156 | 0.079 | 0.115 | 0.153 | 0.192 |
| 2        | 3  | 0.078 | 0.110 | 0.140 | 0.170 | 0.065 | 0.083 | 0.101 | 0.119 | 0.070 | 0.092 | 0.114 | 0.136 | 0.079 | 0.112 | 0.147 | 0.181 | 0.079 | 0.115 | 0.153 | 0.192 |
| 3        | 4  | 0.091 | 0.138 | 0.186 | 0.230 | 0.097 | 0.109 | 0.147 | 0.181 | 0.087 | 0.112 | 0.147 | 0.181 | 0.079 | 0.112 | 0.147 | 0.181 | 0.079 | 0.115 | 0.153 | 0.192 |
| 4        | 5  | 0.103 | 0.165 | 0.227 | 0.287 | 0.107 | 0.165 | 0.227 | 0.287 | 0.111 | 0.178 | 0.242 | 0.300 | 0.103 | 0.165 | 0.227 | 0.287 | 0.111 | 0.178 | 0.242 | 0.300 |
| 5        | 2  | 0.060 | 0.070 | 0.081 | 0.091 | 0.067 | 0.085 | 0.100 | 0.114 | 0.067 | 0.085 | 0.100 | 0.114 | 0.067 | 0.085 | 0.100 | 0.114 | 0.067 | 0.085 | 0.100 | 0.114 |
| 3        | 3  | 0.070 | 0.092 | 0.114 | 0.136 | 0.084 | 0.119 | 0.152 | 0.182 | 0.084 | 0.119 | 0.152 | 0.182 | 0.084 | 0.119 | 0.152 | 0.182 | 0.084 | 0.119 | 0.152 | 0.182 |
| 4        | 4  | 0.098 | 0.150 | 0.199 | 0.244 | 0.111 | 0.178 | 0.242 | 0.300 | 0.111 | 0.178 | 0.242 | 0.300 | 0.111 | 0.178 | 0.242 | 0.300 | 0.111 | 0.178 | 0.242 | 0.300 |
| 5        | 5  | 0.087 | 0.131 | 0.177 | 0.223 | 0.087 | 0.131 | 0.177 | 0.223 | 0.087 | 0.131 | 0.177 | 0.223 | 0.087 | 0.131 | 0.177 | 0.223 | 0.087 | 0.131 | 0.177 | 0.223 |
Table 3 shows the two-sided \( p \) values of the adjusted CMH test with respect to different values of \( C \) and \( D \) for scenarios 1–5. Note that the strength parameter \( C = \exp(\gamma) \) is the odds ratio of a black mother having a LBW baby compared with a white mother. The imbalance parameter \( D = \exp(\gamma) \) is the increased odds ratio of a mother being black if she has hypertension. When \( C = 1 \) or \( D = 1 \), race is not a confounder and the CMH test is unchanged.

Another way to perform the sensitivity analysis is to obtain the threshold values of the sensitivity parameters \( C \) and \( D \), which increase the \( p \) value of the CMH test to 0.05. Figure 1 shows the plots of sensitivity parameters \( C \) and \( D \), which corresponds to \( p = 0.05 \). From figure 1, we see that when \( C = 1.5 \) and \( D = 1.1 \), then the two-sided \( p \) value would reach the 0.05 level. Thus, both table 3 and figure 1 show that the \( p \) value of the CMH test is sensitive to moderate changes in \( C \) and \( D \), indicating that the significant association between maternal hypertension and LBW might have arisen from a hidden bias due to confounding. Hence, other studies, including race and ethnicity information, are needed to reach the conclusion that maternal hypertension causes LBW. If the studies demonstrate that increased incidence of black mothers with hypertension was slight with odds ratio less than 1.5, then the significance of the London study would not be affected by the omission of race.

Although in Rubin’s causal model, sex and race are not used as ‘treatment’ variables in causal inference because it is difficult to ascribe causality to variables that cannot be altered (Holland 1986), here race is used as a proxy for other
factors reflecting socio-economic disparities. More recently, race and sex have been used in the propensity score method to form ‘similar strata’ (Zanutto et al. 2005) but should not be the main potential outcome of a study as they cannot be altered.

(b) Sensitivity analysis of allopurinol as a cause of rash

Rosenbaum (2002) examined the sensitivity of a study indicating that allopurinol can cause rash (Boston Collaborative Drug Project 1972). Allopurinol is commonly used for the treatment and prevention of attacks of gout and certain types of kidney stones. It is also used to treat elevated uric acid levels in the blood and urine, which can occur in patients receiving chemotherapy for the treatment of leukaemia, lymphoma and other types of cancer. Allopurinol is usually well tolerated by most patients; however, some may experience possible side effects of skin rash, hives and itching.

Table 4 shows the joint frequencies of drug use and getting a rash case for males and females, respectively. The CMH test is highly significant with \( p<0.0001 \). In a randomized trial or in the absence of unobserved confounders, this indicates strong evidence that allopurinol causes rash (Rosenbaum 2002, p. 131). Because the patients could also be taking amoxicillin and ampicillin, which are known to cause rash, is it possible that the rash was caused by an unobserved factor, e.g. use of amoxicillin or other drugs, instead of allopurinol?

Although information about the other drug use and users’ allergic history was not collected, one can assess the sensitivity of the CMH test with respect to an unobserved confounder. Let \( U \) be the binary indicator for amoxicillin use, e.g. use amoxicillin (1) and no amoxicillin (0). Again, the prevalences of using amoxicillin among the non-allopurinol users are denoted by \( q_j, j=1,\ldots, J \). The parameter \( \delta \) indicates the imbalance of \( U \) between allopurinol users and non-users. The parameter \( \gamma \) indicates the increasing probability of having rash among users of amoxicillin. We assume two possible scenarios for the prevalences of amoxicillin use among the non-allopurinol users, i.e. \( q_j=10\% \) or \( 20\% \), where \( j=1, 2 \).

The increasing odds ratio of having rash for amoxicillin users compared with non-users is not known, but it is likely to be below 4. Based on table 5, we can see that the CMH test of allopurinol causing rash is not sensitive to the omitted variable (use of amoxicillin). Even when the strength parameter \( C=4 \) and the imbalance parameter \( D=4 \), the CMH test remains highly significant with \( p=0.002 \). This example also illustrates the additional information provided...
by the simultaneous sensitivity analysis compared with the earlier approaches by Cornfield et al. (1959) and Rosenbaum (2002), which focus primarily on the imbalance parameter and assume a confounder with infinitely large strength. Rosenbaum (2002, p. 131) used the parameter $G$, which is similar to $D$ in our analysis, to measure the imbalance of treatment assignment with respect to the unobserved factor. The upper bounds on the $p$ values based on his analysis are 0.036 and 0.30, for $G = 2$ and 3, respectively, which indicates that the causal relationship between allopurinol use and rash could have arisen from an unobserved confounder that triples the risk of rash. While such a confounding is unlikely to exist, the conclusion based on the simultaneous analysis is stronger. Even when $C = \infty$, the test barely reached non-significance for $D = 3$ and $q_j = 10\%$. This sensitivity analysis is more stringent, indicating the conclusion that allopurinol causes rash is very robust to an unobserved confounder. This is consistent with the current medical practice that lists rash as the one common side effect of allopurinol use (http://www.nlm.nih.gov/medlineplus/druginfo/medmaster/a682673.html).

### 4. Discussion

Because the CMH test is widely used to analyse the data obtained in epidemiological studies, legal cases and social sciences, it is important to conduct a sensitivity analysis to answer whether that inference could plausibly be explained by an unobserved confounder before drawing a firm conclusion.

The proposed simultaneous sensitivity analysis approach is similar to the primal sensitivity analysis method by Rosenbaum (2002). The major difference is the introduction of the strength parameter. The primal sensitivity analyses closely parallel the theory of randomized experiments and have been developed for many statistical tests and estimators, e.g. the McNemar and Wilcoxon signed-rank tests and the Hodges–Lehmann estimator (Rosenbaum 2002). The simultaneous approach allows one to use more subject matter knowledge. However, it also requires more assumptions about both relationships. The simultaneous approach is especially useful when an observational study is challenged owing to failure to control for a particular unobserved variable, when there are plausible ranges for the values of $\gamma$ and $\delta$ (Gastwirth et al. 1998).
At the design stage, it is also useful to examine the size and power of the corrected CMH test in the presence of omitted confounders in a range of realistic values of the strength and imbalance parameters. The calculation of size and power for the Cochran–Armitage trend test has been derived by Yu & Gastwirth (2005). A similar technique can be used for the CMH test with omitted variable.

In future studies, we plan to extend this methodology of sensitivity analysis to data where the unobserved confounders are continuous and ordinal. In real-world observational studies, there may be more than one omitted confounder; background information on the joint distribution of the multiple confounders can then be used to replace the univariate distribution (2.5) in a sensitivity analysis.

B.Y. was supported in part by the Intramural Research Programme of the National Institute on Aging and J.L.G. was supported in part by grant SES-0317956 from the National Science Foundation.

Appendix A. Estimation of \((\pi_{1j}, \pi_{0j})\)

This appendix describes the estimation of \((\pi_{1j}, \pi_{0j})\) under the product-binomial likelihood (2.1) with fixed values of \((q_1, \ldots, q_J, \beta, \gamma, \delta)\) for the allopurinol data in §3b. For stratum \(j\), the prevalence of \(U\) in the non-exposed group is \(q_j = \exp(\lambda_j)/[1 + \exp(\lambda_j)]\). Then the prevalence of \(U\) in the exposed group is \(r_j = q_j e^\delta/(1 - q_j + q_j e^\delta)\). Let \(A_j = \exp(\alpha_j), B = \exp(\beta), C = \exp(\gamma)\). Then for fixed values of \((\beta, \delta, \gamma, q_j)\), the marginal probabilities of response are

\[
\pi_{1j} = \frac{A_j BC}{1 + A_j BC} r_j + \frac{A_j B}{1 + A_j B} (1 - r_j)
\]  

(A 1)

and

\[
\pi_{0j} = \frac{A_j C}{1 + A_j C} q_j + \frac{A_j}{1 + A_j} (1 - q_j).
\]  

(A 2)

The log-likelihood of function (2.1) is

\[
\log L(\theta|D) = \sum_{j=1}^{J} \ell_j(\theta|D),
\]

where

\[
\ell_j(\theta|D) = a_j \log \pi_{1j} + b_j \log (1 - \pi_{1j}) + c_j \log \pi_{0j} + d_j \log (1 - \pi_{0j}),
\]

for stratum \(s\). The maximum-likelihood estimate (MLE) of \(\alpha_j\), and \((\pi_{1j}, \pi_{0j})\) could be obtained from SAS proc nlp. The code for the following \(2 \times 2 \times 2\) table from the allopurinol study with \(q_1 = q_2 = 0.2, \beta = 0, \gamma = \log(2)\) and \(\delta = \log(2)\) is shown below. The MLEs of \((\pi_{1j}, \pi_{0j})\) are given in PPIEST.
References

Boston Collaborative Drug Project 1972 Excess of ampicillin rashes associated with allopurinol or hyperuricemia. *N. Engl. J. Med.* 286, 505–507.
Centers for Disease Control and Prevention 2002 Infant mortality and low birth weight among black and white infants—United States, 1980–2000. *Morb. Mortal. Wkly Rep.* 51, 589–592.
Cochran, W. G. 1954 Some methods for strengthening the common X2 tests. *Biometrics* 10, 417–451. (doi:10.2307/3001616)
Cornfield, J., Haenszel, W., Lilienfeld, A., Shimkin, M. & Wynder, E. 1959 Smoking and lung cancer. *J. Natl Cancer Inst.* 22, 173–203.

*Phil. Trans. R. Soc. A* (2008)
David, R. J. & Collins, J. W. 1997 Differing birth weight among infants of U.S. born blacks, African-born blacks, and U.S. born whites. *N. Engl. J. Med.* **337**, 1209–1214. (doi:10.1056/NEJM199710233371706)

Gastwirth, J. L. 1988 *Statistical reasoning in law and public policy*. San Diego, CA: Academic Press.

Gastwirth, J. L. 1992 Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33**, 19–34.

Gastwirth, J. L., Krieger, A. M. & Rosenbaum, P. R. 1998 Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85**, 907–920. (doi:10.1093/biomet/85.4.907)

Gerominus, A. T., Keene, D., Hicken, M. & Bound, J. 2007 Black–white differences in age trajectories of hypertension prevalence among adult women and men, 1999–2002. *Ethn. Dis.* **17**, 40–48.

Hills, M. & De Stavola, B. 2002 *A short introduction to Stata 8 for biostatistics*. London, UK: Timberlake Consultants Ltd. See http://www.timberlake.co.uk.

Holland, P. W. 1986 Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960. (doi:10.2307/2289064)

Mantel, N. & Haenszel, W. 1959 Statistical aspects of retrospective studies of disease. *J. Natl Cancer Inst.* **22**, 719–748.

Rosenbaum, P. R. 2002 *Observational studies*, 2nd edn. New York, NY: Springer.

Rosenbaum, P. R. & Krieger, A. 1990 Sensitivity analysis for two-sample permutation inference in observational studies. *J. Am. Stat. Assoc.* **85**, 493–498. (doi:10.2307/2289789)

Rosenbaum, P. R. & Rubin, D. B. 1983 Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcomes. *J. R. Stat. Soc. Ser. B* **45**, 212–218.

Rosenbaum, P. R. & Rubin, D. B. 1984 Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **79**, 516–524. (doi:10.2307/2288398)

US Department of Health and Human Services, 2000 *Healthy people 2010*, vols. 2 (conference ed.). Washington, DC: US Department of Health and Human Services.

Woolson, R. F., Bean, J. A. & Rojas, P. B. 1986 Sample size for case-control studies using Cochran’s statistic. *Biometrics* **42**, 927–932. (doi:10.2307/2530706)

Yu, B. & Gastwirth, J. L. 2005 Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics* **6**, 201–209. (doi:10.1093/biostatistics/kxi003)

Zanutto, E. L., Lu, B. & Hornik, R. 2005 Using propensity score subclassification for multiple treatment doses to evaluate a national anti-drug media campaign. *J. Educ. Behav. Stat.* **30**, 59–73. (doi:10.3102/107699860300001059)