

Cochran-Mantel-Haenszel Techniques:
Applications Involving Epidemiologic Survey Data

*Daniel B. Hall, Robert F. Woolson, William R. Clarke
and Martha F. Jones*

1. Introduction

In epidemiologic research, the relationship between the presence or absence of exposure and health status is often of interest. Commonly, one or more extraneous variables are present, on which the relationship depends. Such variables may be confounders or effect modifiers. In either case, the marginal relationship between exposure and health status can be quite different than the conditional relationships that exist at particular levels of these variables. This phenomenon is well known as Simpson's paradox. When these conditional relationships differ over the levels of the other variables, it is occasionally best to measure and report these conditional relationships separately. However, it is frequently the case that some measure of the strength of the relationship can be found which can be reasonably assumed to be constant across the levels of the confounding or effect modifying variables. In this case, a summary statistic or test of average partial association is preferable.

For example, in a cross-sectional study on the relationship between exercise and breast cancer, one method for dealing with the potentially confounding effects of age would be to report differences in the prevalence of breast cancer between sedentary and active subjects at each of several age levels. One likely scenario in this setting is that the prevalence of breast cancer increases with age in both the sedentary and active populations but the difference between these prevalences does

not remain constant. In this case it would be inappropriate to combine age-specific differences into a single prevalence difference for all women. It may be the case, though, that the ratio of the prevalences in the sedentary and active groups does stay the same as age increases. If so, it would be desirable and appropriate to combine the age-specific prevalence ratios into a single estimate or test statistic.

In this paper we consider techniques for combining information concerning a two-way association that is conditional on the levels of extraneous variables. That is, we are concerned with measures and tests of two-way association that control for the effects of confounding or effect modifying variables. In particular, we will focus on techniques based on the papers of Cochran (1954) and Mantel and Haenszel (1959). These Cochran-Mantel-Haenszel (CMH) techniques will be illustrated by means of a case study of the Health Assessment of Persian Gulf War Veterans from Iowa (Iowa PGW Study, The Iowa Persian Gulf Study Group, 1997), an epidemiologic investigation based on survey data from a stratified random sample of Iowans. Adaptations of CMH methods will be introduced for stratified sampling designs of the sort used in the Iowa PGW Study and the performance of the proposed methods will be examined in a simulation study.

2. CMH techniques: a review

In Cochran's original (1954) paper, a test statistic was introduced to extend the chi-square test of independence in a 2×2 table to multiple 2×2 tables where each table corresponds to a different level of an intervening variable. That is, Cochran proposed a test of conditional independence — independence of the variables forming the rows and columns of the tables, conditional on the levels of a third variable. To establish notation, let n_{hij} represent the number of responses observed at the i^{th} level of the row variable, the j^{th} level of the column variable, and the h^{th} level of the intervening variable. Assuming H levels of the intervening variable, $I = 2$ rows, and $J = 2$ columns, we have data that may be summarized as in Table 1, for

Table 1.

Exposure Status	Health Status		Total
	Condition Present	Condition Absent	
Not Exposed	n_{h11}	n_{h12}	$n_{h1\cdot}$
Exposed	n_{h21}	n_{h22}	$n_{h2\cdot}$
Total	$n_{h\cdot1}$	$n_{h\cdot2}$	$n_{h\cdot\cdot}$

$h = 1, \dots, H$.

For this situation Cochran conditioned on the row totals, considering each 2×2 table to consist of independent binomials. He based his statistic on a weighted sum of the table-specific differences in proportions, $d_w = \sum_{h=1}^H w_h (\hat{p}_{h1} - \hat{p}_{h2})$, where $w_h = n_{h1} \cdot n_{h2} / n_{h\cdot\cdot}$ and $\hat{p}_{hi} = n_{hi1} / n_{hi\cdot}$, $i = 1, 2$. Using the asymptotic normality of d_w , he justified

$$X_C^2 = \frac{d_w^2}{\text{var}(d_w)} = \frac{d_w^2}{\sum_h n_{h1} \cdot n_{h2} \cdot n_{h\cdot1} n_{h\cdot2} / n_{h\cdot\cdot}^3}$$

as an appropriate test statistic, having an approximate $\chi^2(1)$ distribution under the null hypothesis of conditional independence.

Mantel and Haenszel (1959) proposed a similar test statistic using a hypergeometric assumption. Conditional on the row and column totals, the cell counts in each table have a hypergeometric distribution. This fact suggests a test statistic based on the difference between the observed and expected frequencies in each 2×2 table. As with the classic chi-square test of independence in a single 2×2 table, it suffices to compare the observed and expected count in one cell per table. The Mantel-Haenszel test is, therefore,

$$X_{MH}^2 = \frac{(|\sum_h (n_{h11} - n_{h1} \cdot n_{h\cdot1} / n_{h\cdot\cdot})| - \frac{1}{2})^2}{\sum_h \frac{n_{h1} \cdot n_{h2} \cdot n_{h\cdot1} \cdot n_{h\cdot2}}{n_{h\cdot\cdot}^2 (n_{h\cdot\cdot} - 1)}} = \frac{(|d_w| - \frac{1}{2})^2}{\sum_h w_h n_{h\cdot1} n_{h\cdot2} / n_{h\cdot\cdot} (n_{h\cdot\cdot} - 1)}.$$

Aside from the continuity correction in X_{MH}^2 , the two test statistics differ by a factor of $n_{h\cdot\cdot} / (n_{h\cdot\cdot} - 1)$ in each table. For moderate to large sample sizes per table,

the difference between the two statistics is typically negligible. In general, though, X_{MH}^2 offers advantages. Both statistics are asymptotically $\chi^2(1)$, but the quality of this approximation depends upon the table-specific sample sizes only for X_{C}^2 . In the extreme, Mantel and Haenszel's statistic will perform adequately in matched pair studies in which $n_{h..} = 2$ for all h , for sufficient total sample size. Cochran's statistic should not be used for such a situation. In addition, the Mantel-Haenszel test has been shown to be optimal under the assumption of a constant odds ratio across tables (Birch, 1964), and it is asymptotically equivalent to likelihood ratio tests from unconditional and conditional logistic regression models for large strata and sparse data situations, respectively (Breslow and Day, 1980).

In addition to the test statistic, X_{MH}^2 , Mantel and Haenszel proposed an odds ratio estimator in their original 1959 paper. Their estimator is a weighted average of the table-specific observed odds ratios:

$$\begin{aligned}\hat{\psi}_{\text{MH}} &= \frac{\sum_h v_h \hat{\psi}_h}{\sum_h v_h} \\ &= \frac{\sum_h R_h}{\sum_h S_h},\end{aligned}$$

where $v_h = n_{h12}n_{h21}/n_{h..}$, $\hat{\psi}_h = n_{h11}n_{h22}/n_{h12}n_{h21}$, $R_h = n_{h11}n_{h22}/n_{h..}$, and $S_h = n_{h12}n_{h21}/n_{h..}$.

Since the introduction of X_{C}^2 , X_{MH}^2 , and $\hat{\psi}_{\text{MH}}$, a large literature has developed. Mantel-Haenszel-type estimators have been developed for the rate ratio (Rothman and Boice, 1979), rate difference (Greenland, 1982), risk ratio (Rothman and Boice, 1979; Tarone, 1981; Nurminen, 1981; and Kleinbaum, Kupper, and Morgenstern, 1982), and risk difference (Greenland, 1982). The distributions of these estimators have been considered and comparisons with other estimators have been made in papers by many authors (Hauck, 1979; Breslow, 1981; Breslow and Liang, 1982; Flanders, 1985; Greenland and Robins, 1985; Robins, Breslow, and Greenland, 1986; Sato, 1990; O'Gorman, Woolson, and Jones, 1994; etc.). Generalizations of $\hat{\psi}_{\text{MH}}$ to $H \times J$ tables ($J > 2$) have been introduced by Mickey and Elashoff

(1985); Liang (1987); Greenland (1989); and Yanagawa and Fujii (1990, 1995). Generalizations of Cochran's and Mantel and Haenszel's tests have been discussed by Hopkins and Gross (1971); Birch (1965); Suguira and Otake (1974); Mantel (1963); Landis, Heyman, and Koch (1978); and Sen (1988). Adjustments to CMH tests and estimators to account for complex sampling designs have been discussed by Donald and Donner (1987); Graubard, Fears, and Gail (1989); and Weerasekera and Bennett (1992). A comprehensive review of this literature is beyond the scope of this paper. The interested reader is referred to the review by Kuritz, Landis, and Koch (1988).

Here we will limit attention to CMH techniques which arose during the analysis of the Iowa PGW Study data. The variety of outcomes considered in this study, and the multiplicity of research questions addressed, provide an excellent showcase for the CMH methodology. In addition, the complication provided by the Iowa PGW Study's stratified sampling scheme motivates some interesting and novel adjustments to standard CMH methods.

At the beginning of section 3 we introduce the Iowa PGW Study. Estimation problems in $H \times J$ tables are discussed in subsections 3.1 and 3.2, and hypothesis testing in such tables is dealt with in subsection 3.3. In terms of estimation, relatively little has been written about Mantel-Haenszel methods in the general $H \times I \times J$ contingency table case and more work in this area seems warranted. As for testing, the generalized CMH test statistics of Landis *et al.* (1978) apply to the general three-way case and the methods of section 3.3 can be extended to this case without essential modification. For other generalized CMH tests for $I > 2$ such as Sen's (1988) union-intersection version of the CMH test for restricted alternative hypotheses, modifications to account for a non-simple random sampling design are more complicated and the methods of section 3.3 do not directly apply.

3. The Iowa PGW Study

The Iowa PGW Study was an epidemiologic study in which a sample of 3695 Iowans who served in the military during the time period of the Persian Gulf War (PGW) were surveyed to assess a variety of health outcomes and exposures. The survey was administered using computer assisted telephone interviewing to a sample from a population of 28,968 individuals who met the inclusion criteria. This sample was obtained using stratified random sampling with proportional allocation, where the stratification was done by exposure status (deployed to the PGW military theater versus not deployed to the PGW theater), regular military versus guard/reserve status, branch of service, rank (enlisted versus officer), gender, race (white versus black and other races), and age (≤ 25 years versus > 25). The distributions of the population, intended sample, and achieved sample over these seven stratification variables are available on request from the authors.

The exposure of primary interest in the Iowa PGW Study was deployment to the PGW military theater. With respect to this exposure, the Iowa PGW Study is similar to a retrospective cohort study. Subjects serving in the military during the PGW time frame were sampled on the basis of their exposure status and data concerning subsequent health experiences were collected at the time of the survey administration in 1995-96. However, because most of the questions in the Iowa PGW Study's survey ask about current and recent health experiences rather than all Post-PGW health experiences, the study was not a cohort study. Rather it combined elements of both the cohort and cross-sectional designs. The result was that for the vast majority of the health outcomes considered, the Iowa PGW Study allowed for the estimation of prevalence but not incidence.

The hypotheses of interest in the Iowa PGW Study can be described in terms of the four population domains described in Table 2.

The hypothesis of primary interest was

H₁ The current health status of military personnel who were deployed to the PGW

Table 2. Iowa PGW Study Domains.

	Exposed	Not Exposed
Regular Military	Domain 1	Domain 2
National Guard/Reserve	Domain 3	Domain 4

theater (domains 1 & 3) is no different than that of military personnel serving at the time of the PGW who were not deployed to the PGW theater (domains 2 & 4).

There were also three hypotheses of secondary interest:

H₂ Among regular military personnel, the current health status of those deployed to the PGW theater (domain 1) is no different than for those who served during the period of the PGW but were not deployed in theater (domain 2).

H₃ Among National Guard/Reserve military personnel, the current health status of those deployed to the PGW theater (domain 3) is no different than for those who were activated during the period of the PGW but were not deployed in theater (domain 4).

H₄ Among military personnel deployed to the PGW theater, the current health status of those in the regular military (domain 1) is no different than for those serving in the National Guard/Reserve (domain 3).

The primary goals of the initial statistical analysis of Iowa PGW Study data were as follows:

- Test the hypotheses of interest (**H₁**, **H₂**, **H₃**, **H₄**);
- Obtain point estimates and standard errors (SEs) of quantities which summarize the comparisons implicit in the four study hypotheses;
- Obtain point estimates and SEs of prevalence and, where possible, incidence within the four study domains.

In addition, there were several secondary goals of the initial statistical analysis, most of which pertained to study methodology. An overview of the analytical methods that were used to meet these goals may be found in Jones *et al.* (1998). We will limit discussion here to the first two primary goals. For most of the health outcomes measured from Iowa PGW Study data, CMH techniques provided the statistical tests of $\mathbf{H}_1, \dots, \mathbf{H}_4$ and the estimators corresponding to the comparisons implicit in $\mathbf{H}_1, \dots, \mathbf{H}_4$.

3.1 Estimation – dichotomous outcomes

Most common among the Iowa PGW Study outcomes were dichotomous responses indicating the presence or absence of various adverse health conditions. For example, the presence or absence of major depression, bronchitis, and cognitive dysfunction were all measured based on the self-reported symptoms of study respondents. For these dichotomous outcomes (as for all study outcomes), comparisons of interest were made controlling for the stratification variables age, sex, race, rank, branch of service, and, for \mathbf{H}_1 , National Guard/Reserve status. In addition, there were a few outcomes for which covariates other than the stratification variables were controlled in the statistical analysis. For example, smoking status (current smoker, former smoker, non-smoker) was controlled in the analysis of respiratory outcomes such as bronchitis. The levels of these extraneous variables (stratification variables and, possibly, covariates) can be thought of as separating the data into H 2×2 tables of the form given in Table 1. Here, H is the number of combinations of the levels of the extraneous variables. Although the separation of the data into H tables does correspond to “stratification” in some sense, we will not use this terminology to avoid confusion with the stratification of the Iowa PGW Study’s sampling design.

For these dichotomous outcomes the basic Mantel-Haenszel (1959) test described in section 1 is appropriate, provided that adjustments for the stratified sampling

design are made. Since this test is a special case of the Generalized Cochran-Mantel-Haenszel (GCMH) test to be described in section 3.2, we will postpone discussion of such adjustments until then. Besides testing, though, it was of interest to quantify the comparisons inherent in the study hypotheses, $\mathbf{H}_1, \dots, \mathbf{H}_4$. That is, for each dichotomous outcome, a measure of average partial association was desired for data described by H 2×2 tables with columns corresponding to presence and absence of the outcome and rows corresponding to exposure status, in the case of hypotheses $\mathbf{H}_1, \mathbf{H}_2$, and \mathbf{H}_3 , or National Guard/Reserve status, in the case of hypothesis \mathbf{H}_4 . The particular measure of association that was of primary interest was the prevalence difference. This quantity was chosen *a priori* because, unlike measures of association based on ratios, the prevalence difference provides a direct measure of the public health impact of the exposure (in this case, the Persian Gulf War). In addition, though, the odds ratio and (prevalence) rate ratio were also of interest for some outcomes. In the remainder of this section we will consider the Mantel-Haenszel estimators for these quantities and discuss adjustments appropriate to the stratified sampling design of the Iowa PGW Study.

Ignoring for the time being the sampling design of the Iowa PGW Study, one approach to modelling the Iowa PGW data is to think of the responses for any given health outcome as arising from H pairs of independent binomial variables (n_{h11}, n_{h21}) , with denominators $(n_{h1\cdot}, n_{h2\cdot})$, and probabilities of disease (p_{h1}, p_{h2}) $h = 1, \dots, H$. Such an approach is generally appropriate for data from fixed cohort studies of exposure and disease. An alternative assumption appropriate in some situations is that (n_{h11}, n_{h21}) , $h = 1, \dots, H$, is a series of independent Poisson random variables with fixed “person-time” denominators $(n_{h1\cdot}, n_{h2\cdot})$ and rates (r_{h1}, r_{h2}) , $h = 1, \dots, H$. Such an assumption would fit a follow-up study of a dynamic population where loss to follow-up is expected. Under the binomial assumption the parameters of interest are probabilities, or risks. Under the Poisson assumption, the parameters of interest are rates.

In the binomial situation, measures of association between exposure and disease include the risk difference and risk ratio. For the h^{th} pair of observations (n_{h11}, n_{h21}) , the risk difference, δ_h , is defined as $\delta_h = p_{h1} - p_{h2}$. When this difference is assumed constant across h , it makes sense to estimate the common risk difference δ as a weighted average of the estimated δ_h , $h = 1, \dots, H$. Using Cochran's weights, $w_h = n_{h1}n_{h2}./n_{h..}$, $h = 1, \dots, H$, yields the Mantel-Haenszel risk difference estimator (Greenland, 1982),

$$\hat{\delta}_{\text{MH}} = \frac{\sum_h w_h(\hat{p}_{h1} - \hat{p}_{h2})}{\sum_h w_h} = \frac{\sum_h (n_{h11}n_{h2}./n_{h..} - n_{h21}n_{h1}./n_{h..})}{\sum_h n_{h1}n_{h2}./n_{h..}}.$$

When interest centers on the relative risk in the exposed versus control groups, the risk ratio is a more appropriate parameter. The risk ratio for the h^{th} pair (n_{h11}, n_{h21}) , is $\phi_h = p_{h1}/p_{h2}$. Under the constant risk ratio assumption that $\phi_h = \phi$, $h = 1, \dots, H$, a weighted average of table-specific estimates ($\hat{\phi}_h$, $h = 1, \dots, H$) is, again, an appropriate estimator. Weighting by an estimate of the reciprocal asymptotic variance leads to the Mantel-Haenszel risk ratio estimator (Rothman and Boice, 1979; Tarone, 1981),

$$\hat{\phi}_{\text{MH}} = \frac{\sum_h n_{h11}n_{h2}./n_{h..}}{\sum_h n_{h21}n_{h1}./n_{h..}}.$$

Although several alternatives to the Mantel-Haenszel risk ratio estimator have been considered in the literature, only $\hat{\phi}_{\text{MH}}$ is dually consistent (Greenland and Robins, 1985). That is, only $\hat{\phi}_{\text{MH}}$ is consistent under the "large-stratum" assumption that $n_{h1.}$, and $n_{h2.}$ tend to infinity for all h , and under "sparse-data" asymptotics (Breslow, 1981). In the sparse-data model, as the total sample size increases the number of tables increases, but the number of possible denominator configurations $(n_{h1.}, n_{h2.})$ that may occur is assumed to be finite. While the conditional maximum likelihood estimator is also consistent in this case, the unconditional maximum likelihood estimator (Rothman and Boice, 1979), weighted least squares estimator (Grizzle, Starmer, and Koch, 1969; Greenland and Robins,

1985), Rothman and Boice's (1979) "null-weighted" least squares estimator, and Tarone's estimator ϕ_1 are not, and are consistent only under the large-stratum asymptotic situation.

Under the Poisson model, measures of association between exposure and disease include the rate difference and rate ratio. For the h^{th} 2×2 table, let $\gamma_h = r_{h1} - r_{h2}$ and $\omega_h = r_{h1}/r_{h2}$ be the rate difference and ratio, respectively. Assuming $\gamma_h = \gamma$, $h = 1, \dots, H$, or $\omega_h = \omega$, $h = 1, \dots, H$, we have Mantel-Haenszel estimators $\hat{\gamma}_{\text{MH}} = \hat{\delta}_{\text{MH}}$ (Greenland, 1982), or $\hat{\omega}_{\text{MH}} = \hat{\phi}_{\text{MH}}$ (Rothman and Boice, 1979).

Under either the binomial or Poisson model, the odds ratio may be the preferred measure of association between exposure and disease. In case-control studies, neither the risk ratio nor the rate ratio can be estimated because the column totals, $n_{h \cdot 1}$, $n_{h \cdot 2}$, are fixed. Instead ϕ or ω is typically estimated by the odds ratio, ψ , under a rare disease assumption. Alternatively, the odds ratio may be of interest in and of itself. Under the constant odds ratio assumption that $\psi = p_{h1}(1 - p_{h2})/p_{h2}(1 - p_{h1})$, for all h , the Mantel-Haenszel odds ratio estimator (Mantel and Haenszel, 1959) is

$$\hat{\psi}_{\text{MH}} = \frac{\sum_h R_h}{\sum_h S_h},$$

given in section 1. Breslow (1981) has demonstrated the consistency of $\hat{\psi}_{\text{MH}}$ under a sparse-data asymptotic model. As noted earlier, the conditional maximum likelihood odds ratio estimator is also consistent under this model, but $\hat{\psi}_{\text{MH}}$ has a computational advantage without sacrificing much efficiency. Other odds ratio estimators such as maximum likelihood and empirical logit are consistent only under the large-stratum model.

Several estimators of the variance of $\hat{\psi}_{\text{MH}}$ have been introduced (Breslow, 1981; Breslow and Liang, 1982; Connett, Ejigou, McHugh, and Breslow, 1982; Flanders, 1985; Fleiss, 1984; Gilbaud, 1983; Hauck, 1979; Phillips and Holland, 1987; Robins *et al.*, 1986; Ury, 1982). These estimators are reviewed in Kuritz *et al.* (1988). Although none of these estimators has been shown to be "best" according to objective

criteria, Kuritz *et al.* conclude that the Robins, Breslow, and Greenland (1986) estimator and the similar Flanders (1985) estimator are the “formulae of choice” (Kuritz *et al.*, 1988, p.134). These variance estimators are consistent under both the large-stratum and sparse-data asymptotic models, and offer computational convenience. The Robins *et al.* estimator is

$$\hat{\text{var}}_{\text{RBG}}(\hat{\psi}_{\text{MH}}) = \frac{(\hat{\psi}_{\text{MH}})^2}{2} \left[\frac{\sum_h P_h R_h}{(\sum_h R_h)^2} + \frac{\sum_h (P_h S_h + Q_h R_h)}{\sum_h R_h \sum_h S_h} + \frac{\sum_h Q_h S_h}{(\sum_h S_h)^2} \right], \quad (1)$$

where $P_h = (n_{h11} + n_{h22})/n_{h..}$ and $Q_h = (n_{h12} + n_{h21})/n_{h...}$. Because of the skewness of the distribution of $\hat{\psi}_{\text{MH}}$, confidence intervals are typically based on $\log(\hat{\psi}_{\text{MH}})$. The estimator of $\text{var}(\log(\hat{\psi}_{\text{MH}}))$ corresponding to (1) is

$$\hat{\text{var}}_{\text{RBG}}(\log(\hat{\psi}_{\text{MH}})) = \hat{\text{var}}_{\text{RBG}}(\hat{\psi}_{\text{MH}})/(\hat{\psi}_{\text{MH}})^2.$$

Using the same approach to variance estimation as Robins *et al.* (1985), Greenland and Robins (1985) proposed variance estimators for $\hat{\delta}_{\text{MH}}$, $\log(\hat{\phi}_{\text{MH}})$, $\hat{\gamma}_{\text{MH}}$, and $\log(\hat{\omega}_{\text{MH}})$.

Under the stratified random sampling design of the Iowa PGW Study, a more appropriate model than the Poisson or binomial models discussed above is a hypergeometric model. Because the population from which the study’s sample was drawn is finite, we can consider each 2×2 table, $(n_{h11}, n_{h12}, n_{h21}, n_{h22})$, as having a population analogue, $(N_{h11}, N_{h12}, N_{h21}, N_{h22})$. Then (n_{h11}, n_{h21}) can be thought of as a pair of independent hypergeometric random variables with parameters $(N_{h11}, N_{h1.}, n_{h1.})$ and $(N_{h21}, N_{h2.}, n_{h2.})$, respectively. Under such an assumption we desire estimators and standard errors for the common risk ratio $\phi = P_{h1}/P_{h2}$, risk difference $\delta = P_{h1} - P_{h2}$ and odds ratio $\psi = P_{h1}(1 - P_{h2})/P_{h2}(1 - P_{h1})$, $h = 1, \dots, H$, where $P_{hi} = N_{hi1}/N_{hi.}$, $i = 1, 2$. These quantities may be estimated using the Mantel-Haenszel estimators $\hat{\phi}_{\text{MH}}$, $\hat{\delta}_{\text{MH}}$, and $\hat{\psi}_{\text{MH}}$, presented earlier in this section. The variance of these estimators, however, is different under this hypergeometric model than under the binomial or Poisson models.

Consider first the Mantel-Haenszel odds ratio estimator, $\hat{\psi}_{\text{MH}}$. As indicated by Robins *et al.* (1985), the asymptotic variance of this estimator is given by

$$\lim_{H \rightarrow \infty} H \text{var}(\hat{\psi}_{\text{MH}}) = \frac{\lim_{H \rightarrow \infty} \sum_h \text{var}(R_h - \psi S_h)/H}{[\lim_{H \rightarrow \infty} \sum_h E(S_h)/H]^2}. \quad (2)$$

These authors considered an estimator of this asymptotic variance of the form

$$H \bar{\text{v}}\text{ar}(\hat{\psi}_{\text{MH}}) = \frac{\sum_h \hat{v}_h(\hat{\psi}_{\text{MH}})/H}{(\sum_h S_h/H)^2},$$

where $\hat{v}_h(\psi)$ is an unbiased estimator of $\text{var}(R_h - \psi S_h)$. Since $\bar{\text{v}}\text{ar}(\hat{\psi}_{\text{MH}})$ is not invariant under interchange of rows in each 2×2 table, Robins *et al.* (1985) proposed their variance estimator as a symmetrized version of $\bar{\text{v}}\text{ar}(\hat{\psi}_{\text{MH}})$:

$$\hat{\text{v}}\text{ar}_{\text{RBG}}(\hat{\psi}_{\text{MH}}) = \frac{1}{2} \left\{ \bar{\text{v}}\text{ar}^{(1)}(\hat{\psi}_{\text{MH}}) + \bar{\text{v}}\text{ar}^{(2)}(\hat{\psi}_{\text{MH}}) \right\},$$

where $\bar{\text{v}}\text{ar}^{(1)}$ is computed on the original table and $\bar{\text{v}}\text{ar}^{(2)}$ is computed after switching rows in the original table. The corresponding variance estimator for $\log(\hat{\psi}_{\text{MH}})$ is computed similarly.

To obtain a variance estimator for $\log(\hat{\psi}_{\text{MH}})$ under the hypergeometric model, we take the same approach. For $h = 1, \dots, H$, $i = 1, 2$, let $Q_{hi} = 1 - P_{hi} = N_{hi2}/N_{hi}$ and let $f_{hi} = (N_{hi} - n_{hi})/(N_{hi} - 1)$ be the finite population correction for row i . Under the hypergeometric model, $\text{var}(R_h - \psi S_h)$ equals

$$\frac{\psi n_{h1} \cdot n_{h2} \cdot}{n_{h\cdot}^2} [f_{h2} n_{h1} \cdot P_{h1} Q_{h1} + f_{h1} f_{h2} (P_{h1} + P_{h2})^2 + f_{h1} n_{h2} \cdot P_{h2} Q_{h2}].$$

Plugging in unbiased estimators p_{hi} , $p_{hi}(n_{hi1} - f_{hi})/(n_{hi} - f_{hi})$ for P_{hi} , P_{hi}^2 , respectively, $i = 1, 2$, we obtain an unbiased estimator, $\hat{\text{v}}\text{ar}(R_h - \psi S_h)$. Dividing by $(\sum_h S_h)^2$ times the Mantel-Haenszel estimator and symmetrizing we obtain an estimator of $\text{var}(\log(\hat{\psi}_{\text{MH}}))$ under stratified random sampling:

$$\begin{aligned} \hat{\text{v}}\text{ar}(\log(\hat{\psi}_{\text{MH}})) = & \\ & \sum_h \frac{n_{h1} \cdot n_{h2} \cdot}{2n_{h\cdot}} \left\{ (f_{h1} - n_{h1})^{-1} [f_{h1} f_{h2} (f_{h1} - n_{h1} \cdot (q_{h1}^2 + p_{h1}^2)) - f_{h2} n_{h1}^2 \cdot q_{h1}] \right. \\ & + (f_{h2} - n_{h2})^{-1} [f_{h1} f_{h2} (f_{h2} - n_{h2} \cdot (q_{h2}^2 + p_{h2}^2)) - f_{h1} n_{h2}^2 \cdot q_{h2}] \\ & \left. - 2f_{h1} f_{h2} (q_{h1} - p_{h2} + 2p_{h1} p_{h2}) \right\} / R_+ S_+, \end{aligned}$$

where $S_+ = \sum_h S_h$ and $R_+ = \sum_h R_h$.

If we redefine $R_h = n_{h11}n_{h2\cdot}/n_{h\cdot\cdot}$ and $S_h = n_{h21}n_{h1\cdot}/n_{h\cdot\cdot}$, the Mantel-Haenszel risk ratio estimator has the same form as $\hat{\psi}_{\text{MH}}$: $\hat{\phi}_{\text{MH}} = R_+/S_+$. Under these redefinitions, the asymptotic variance of $\hat{\psi}_{\text{MH}}$ has the same form as $\text{v}\hat{\text{a}}\text{r}(\hat{\psi}_{\text{MH}})$, given in (2). To obtain a variance estimator for $\hat{\phi}_{\text{MH}}$ we notice that

$$\begin{aligned}\text{var}(R_h - \phi S_h) &= \frac{n_{h1\cdot}n_{h2\cdot}}{n_{h\cdot\cdot}} [n_{h2\cdot}P_{h1}Q_{h1}f_{h1} + n_{h1\cdot}P_{h2}Q_{h2}f_{h2}\phi^2] \\ &= \frac{n_{h1\cdot}n_{h2\cdot}\phi}{n_{h\cdot\cdot}} [P_{h2}(n_{h2\cdot}f_{h1} + n_{h1\cdot}f_{h2}\phi) \\ &\quad - P_{h1}P_{h2}(n_{h2\cdot}f_{h1} + n_{h1\cdot}f_{h2})].\end{aligned}$$

Plugging in estimators $\hat{P}_{hi} = p_{hi}$, $i = 1, 2$, and $\hat{\phi} = p_{h1}/p_{h2}$ inside the square brackets we obtain an estimator $\text{v}\hat{\text{a}}\text{r}(R_h - \phi S_h) = \phi D_h$, where

$$D_h = \frac{n_{h1\cdot}n_{h2\cdot}}{n_{h\cdot\cdot}}(n_{h21}f_{h1} + n_{h11}f_{h2}) - \frac{n_{h11}n_{h21}}{n_{h\cdot\cdot}}(n_{h2\cdot}f_{h1} + n_{h1\cdot}f_{h2}).$$

This leads to the proposed estimator for $\text{var}(\log(\hat{\phi}_{\text{MH}}))$ under stratified sampling, $\text{v}\hat{\text{a}}\text{r}(\log(\hat{\phi}_{\text{MH}})) = D_+/S_+R_+$, where D_+ is defined analogously to R_+ and S_+ .

Finally consider the Mantel-Haenszel risk difference, $\hat{\delta}_{\text{MH}}$. It is easily seen that

$$\text{var}(\hat{\delta}_{\text{MH}}) = \frac{\sum_h w_h^2 [\text{var}(\hat{p}_{h1}) + \text{var}(\hat{p}_{h2})]}{(\sum_h w_h)^2}.$$

Plugging in unbiased estimators of $\text{var}(\hat{p}_{hi})$, $i = 1, 2$ we obtain

$$\text{v}\hat{\text{a}}\text{r}(\hat{\delta}_{\text{MH}}) = \frac{\sum_h w_h^2 \left[\frac{p_{h1}q_{h1}}{n_{h1\cdot}-1} \left(\frac{N_{h1\cdot}-n_{h1\cdot}}{N_{h1\cdot}} \right) + \frac{p_{h2}q_{h2}}{n_{h2\cdot}-1} \left(\frac{N_{h2\cdot}-n_{h2\cdot}}{N_{h2\cdot}} \right) \right]}{(\sum_h w_h)^2},$$

as a variance estimator under the hypergeometric model.

3.2 Estimation – polytomous outcomes

In addition to dichotomous responses, several polytomous outcomes with ordered response categories were defined from Iowa PGW Study data. Examples of such outcomes include Reported Health Transition (subject feels that his/her health

has become much worse, somewhat worse, about the same, somewhat better, or much better during the last year) and Spectrum of Injury Severity (injury was present, required medical attention, or required hospitalization). In the case of the outcome Time to Conceive, the covariate Number of Pregnancies was used in addition to the stratification variables (age, sex, race, etc.) to separate the data into $H \times J$ tables. For all other polytomous responses, though, covariates other than the stratification variables were not controlled in the preliminary statistical analysis.

Regardless of whether or not covariates were included among the intervening variables, data for each polytomous outcome were separated into $H \times J$ tables of the form given in table 3.

Table 3.

	Severity Level 1	Severity Level 2	⋯	Severity Level J	Total
Group 1	n_{h11}	n_{h12}	⋯	n_{h1J}	$n_{h1\cdot}$
Group 2	n_{h21}	n_{h22}	⋯	n_{h2J}	$n_{h2\cdot}$
Total	$n_{h\cdot 1}$	$n_{h\cdot 2}$	⋯	$n_{h\cdot J}$	$n_{h\cdot\cdot}$

In this case a multivariate version of $\hat{\delta}_{MH}$ may be defined as

$$\hat{\delta}_{MH} = \mathbf{A} \frac{\sum_h w_h [\hat{\mathbf{p}}_{h1} - \hat{\mathbf{p}}_{h2}]}{\sum_h w_h},$$

where $\hat{\mathbf{p}}_{hi} = n_{hi\cdot}^{-1}(n_{hi1}, n_{hi2}, \dots, n_{hiJ})^T$, $i = 1, 2$ are $J \times 1$ vectors, $\mathbf{A} = (\mathbf{I}_{J-1} : \mathbf{0}_{J-1})$, \mathbf{I}_a is the $a \times a$ identity matrix, and $\mathbf{0}_a$ is the $a \times 1$ vector of zeros. An estimate of the $J - 1 \times J - 1$ variance matrix of $\hat{\delta}_{MH}$ is given by

$$\hat{\text{var}}(\hat{\delta}_{MH}) = \mathbf{A} \frac{\sum_h w_h^2 \sum_{i=1}^2 \frac{N_{hi\cdot} - n_{hi\cdot}}{N_{hi\cdot}(n_{hi\cdot} - 1)} (\text{diag}(\hat{\mathbf{p}}_{hi}) - \hat{\mathbf{p}}_{hi} \hat{\mathbf{p}}_{hi}^T)}{(\sum_h w_h)^2} \mathbf{A}^T,$$

where $\text{diag}(\hat{\mathbf{p}}_{hi})$ denotes the matrix with the vector $\hat{\mathbf{p}}_{hi}$ on the diagonal and zeros elsewhere. For all of the ordered polytomous outcomes in the Iowa PGW Study $\hat{\boldsymbol{\delta}}_{\text{MH}}$, standard errors for the elements of $\hat{\boldsymbol{\delta}}_{\text{MH}}$ given by the square roots of the diagonal elements of $\hat{\text{var}}(\hat{\boldsymbol{\delta}}_{\text{MH}})$, and Scheffé-adjusted 95% confidence intervals for the population quantities corresponding to the elements of $\hat{\boldsymbol{\delta}}_{\text{MH}}$ were reported. No generalized risk ratios or odds ratios were considered for polytomous variables.

3.3 Hypothesis testing

As mentioned above, Mantel and Haenszel's original (1959) test of conditional independence in H 2×2 tables is applicable for testing hypotheses H_1, \dots, H_4 for dichotomous health outcomes, provided that adjustments for the stratified random sampling design of the Iowa PGW Study are made. For polytomous outcomes in H $2 \times J$ tables several generalized Cochran-Mantel-Haenszel (GCMH) tests are available (Landis *et al.*, 1978). Two such GCMH tests, the general association test and the mean score test, were used in the analysis of Iowa PGW data. Both of these tests were performed for ordered polytomous health outcomes. For dichotomous outcomes each of these tests reduces to the Mantel-Haenszel (1959) test. A detailed description of these testing procedures in the context of simple random sampling from an infinite population is provided in Landis *et al.* (1978). Because the PGW Study does not utilize a simple random sampling design modifications to these tests as described by Landis, *et al.* (1978) are necessary. These modifications are described in the remainder of this section.

Let x_{GCMH}^2 denote the GCMH test statistic based on sample data. This statistic is defined as follows:

$$x_{\text{GCMH}}^2 = \mathbf{g}^T \mathbf{v}_{\mathbf{g}}^{-1} \mathbf{g},$$

where $\mathbf{g} = \mathbf{B} \sum_h \mathbf{g}_h(\mathbf{n}_h)$, $\mathbf{g}_h(\mathbf{n}_h) = \mathbf{n}_h - \mathbf{m}_h$, $\mathbf{n}_h = (n_{h11}, \dots, n_{h1J}, n_{h21}, \dots, n_{h2J})^T$, $\mathbf{m}_h = \text{E}(\mathbf{n}_h | H_0) = n_{h\cdot\cdot}(\mathbf{p}_{h*} \otimes \mathbf{p}_{h*})$, $\mathbf{p}_{h*} = n_{h\cdot\cdot}^{-1} \mathbf{n}_{h*} = n_{h\cdot\cdot}^{-1}(n_{h1\cdot}, n_{h2\cdot})^T$,

$p_{h\cdot*} = n_{h\cdot}^{-1}n_{h\cdot*} = n_{h\cdot}^{-1}(n_{h\cdot 1}, \dots, n_{h\cdot J})^T$, and

$$\begin{aligned} \mathbf{v}_{\mathbf{g}} &= \text{var}(\mathbf{g}|H_0) = \mathbf{B} \left(\sum_h \text{var}(\mathbf{n}_h|H_0) \right) \mathbf{B}^T \\ &= \mathbf{B} \left(\sum_h \frac{n_{h\cdot}^2}{n_{h\cdot} - 1} \{[\text{diag}(\mathbf{p}_{h\cdot*}) - \mathbf{p}_{h\cdot*}\mathbf{p}_{h\cdot*}^T] \otimes [\text{diag}(\mathbf{p}_{h\cdot*}) - \mathbf{p}_{h\cdot*}\mathbf{p}_{h\cdot*}^T]\} \right) \mathbf{B}^T. \end{aligned}$$

Let X_{GCMH}^2 denote the population analogue of x_{GCMH}^2 defined similarly in terms of the estimated population counts, $\hat{\mathbf{N}}_h = (\hat{\mathbf{N}}_{h11}, \dots, \hat{\mathbf{N}}_{h2J})^T$, rather than the vector of sample counts, \mathbf{n}_h . The GCMH general association statistic results when $\mathbf{B} = \mathbf{A}$. The GCMH mean score test results when $\mathbf{B} = (a_1, \dots, a_J)^T$, a 1 x J vector of column scores. The null hypothesis for both of these tests is that the frequency counts in the two rows of the h^{th} table can be regarded as a pair of simple random samples of sizes $(n_{h1\cdot}, n_{h2\cdot})$ from a population corresponding to the distribution of the column totals. The general association test is directed at general association alternatives – alternatives under which the row distributions differ in nonspecific patterns. The mean score test is directed toward alternatives under which a measure of location, the weighted mean of assigned column scores, differs across rows.

Under general non-simple random sampling schemes a common method for obtaining an estimate of the variance of a non-linear statistic such as \mathbf{G} or \mathbf{g} is the Taylor linearization procedure (Woodruff, 1971). This procedure was used to compute $\hat{\text{var}}(\mathbf{G}|H_0)$ leading to an adjusted version of X_{GCMH}^2 appropriate for Iowa PGW Study data. The Taylor linearization procedure for the CMH statistic when $J=2$ is described in detail in Weerasekera and Bennett (1992). However, the procedure described here differs from that of Weerasekera and Bennett in that these authors apply it to estimate $\mathbf{v}_{\mathbf{g}}$ and use x_{GCMH}^2 as their test statistic rather than X_{GCMH}^2 . Our usage of X_{GCMH}^2 follows the approach used in SUDAAN (Shah, Folsom, LaVange, Wheelless, Boyle, and Williams, 1995), although SUDAAN implements only the general association version of the test.

The Taylor linearization procedure consists of forming a variable Z for each individual, $i = 1, \dots, n_{h\cdot}$, and each table, $h = 1, \dots, H$, which is a linearized version of the statistic G , and then computing the variance of $\sum_h \sum_{i=1}^{n_{h\cdot}} Z_{hi}$, taking into account the sampling design. That is, for the i^{th} individual falling in the h^{th} table,

$$Z_{hi} = \begin{cases} \frac{\partial \mathbf{G}}{\partial \tilde{N}_{h11}}, & \text{if the individual belongs in the } (1, 1)^{\text{th}} \text{ cell,} \\ \frac{\partial \mathbf{G}}{\partial \tilde{N}_{h12}}, & \text{if the individual belongs in the } (1, 2)^{\text{th}} \text{ cell,} \\ \vdots & \\ \frac{\partial \mathbf{G}}{\partial \tilde{N}_{h2J}}, & \text{if the individual belongs in the } (2, J)^{\text{th}} \text{ cell.} \end{cases}$$

Then $\text{var}(\mathbf{G}|H_0) \approx \hat{\text{var}}_{\text{sd}}(\sum_h \sum_{i=1}^{n_{h\cdot}} Z_{hi})$, where the subscript sd denotes that this variance is to be taken with respect to the sampling design. In the case of the PGW Study where we have a stratified random sampling design from a finite population,

$$\hat{\text{var}}_{\text{sd}}\left(\sum_h \sum_{i=1}^{n_{h\cdot}} Z_{hi}\right) = \frac{n^2}{N^2} \sum_{s=1}^S \frac{N_s}{n_s} (N_s - n_s) \hat{\sigma}_s^2,$$

where s indexes strata from 1 to S , N_s is the population size in stratum s , n_s is the sample size in stratum s , $N = \sum_s N_s$, $n = \sum_s n_s$, and $\hat{\sigma}_s^2$ is the sample variance of Z in stratum s . By “strata” here, it is meant the strata by which the sample was drawn which, in general, do not correspond to the tables indexed by h . In the Iowa PGW Study, the strata usually (in the cases when no covariates were considered) corresponded to the rows of the H $2 \times J$ tables.

Under the null hypothesis, X_{GCMH}^2 is distributed as a chi-square statistic with $J - 1$ degrees of freedom. Following the procedure implemented in SUDAAN, the significance level for the GCMH test is based on transforming X_{GCMH}^2 to an F -statistic with $J - 1$ numerator degrees of freedom and $n - S$ denominator degrees of freedom (see Shah *et al.*, 1995, for more details).

4. Simulation study

To assess the performance of the variance estimators proposed here, a simulation study was conducted based on the general features of the Iowa PGW Study’s sample

and the corresponding population. Several data sets were generated with the same size and demographic make-up as in the Iowa PGW Study population, but with known parameters of association between exposure and disease. That is, for both the odds ratio and rate ratio, seven data sets were generated with known parameters near 0.2, 0.5, 0.67, 1.0, 1.5, 2.0 and 5.0, and for the rate difference four data sets were generated with known parameters near 0.0, 0.05, 0.10, and 0.25. Each of these simulated population data sets contained 28,968 observations with the same distribution with respect to the stratification variables, age, sex, race, rank, branch of service, exposure status and National Guard/Reserve status, as observed in the Iowa population. Within each of the strata defined by the combinations of the levels of age, sex, race, rank, branch of service, and National Guard/Reserve status, the probability of disease in the control (unexposed) group was randomly assigned one of the values 0.05, 0.10, 0.15, 0.20, or 0.25, and the probability of disease in the exposed group was assigned according to the known association parameter. In this way constant association parameters were achieved without assuming the unrealistic scenario of homogeneous prevalences across strata.

From each of these population data sets 2000 stratified random samples with allocation proportional to strata sizes were drawn. Each of these stratified random samples was of size 3000 (750 per domain) in the first set of simulations, and of size 5000 (1250 per domain) in the second set. The actual Iowa PGW Study's sample was designed to be 750 per domain based on sample size calculations made during the development of the study protocol. The actual sample size was somewhat larger than this figure (3695). Each sample was used to compute a Mantel-Haenszel estimate and a pair of corresponding standard errors, one computed using the appropriate stratified sampling variance estimator proposed in section (3.1), and one computed using the standard simple random sampling variance estimator ($\hat{\text{var}}_{\text{RBG}}(\log(\hat{\psi}_{\text{MH}}))$) for the odds ratio and the Greenland and Robins (1985) variance estimators for the risk ratio and risk difference). Estimates

and standard errors were computed based on H 2×2 tables with rows corresponding to unexposed and exposed and columns corresponding to presence and absence of disease. Potentially, in each sample H could be as large as 95, the number of combinations of the levels of age, sex, race, rank, branch of service and National Guard/Reserve status containing at least one population member. In any given sample, though, H was typically substantially smaller than 95 due to sampling zeros.

Results of these simulations are presented in Tables 4 and 5. In these tables $\bar{M}\hat{H}$ refers to the average of the 2000 observed Mantel-Haenszel estimators ($\hat{\psi}_{MH}, \hat{\phi}_{MH}$, or $\hat{\delta}_{MH}$); $s_{\bar{M}\hat{H}}$ denotes the standard deviation of the 2000 observed log Mantel-Haenszel estimators for ψ and ϕ and the 2000 untransformed Mantel-Haenszel estimators for δ ; $\bar{S}\hat{E}$ refers to the average of the 2000 standard errors (for $\log(\hat{\psi}_{MH})$, $\log(\hat{\phi}_{MH})$, or $\hat{\delta}_{MH}$); $s_{S\hat{E}}$ refers to the standard deviation of these standard errors; $\hat{M}\hat{S}\hat{E} = (\bar{S}\hat{E} - s_{\bar{M}\hat{H}})^2 + s_{S\hat{E}}^2$; and coverage refers to the observed percentage of nominal 95% confidence intervals for which the parameter value was covered. Intervals were based on the asymptotic normality of $\log(\hat{\psi}_{MH})$, $\log(\hat{\phi}_{MH})$ and $\hat{\delta}_{MH}$.

From the tables it is apparent that the proposed variance estimator for $\hat{\phi}_{MH}$ has smaller bias and variance than the simple random sampling (SRS) variance estimator in all cases considered. That is, for the risk ratio $\bar{S}\hat{E}$ is consistently closer to $s_{\bar{M}\hat{H}}$ using the stratified sampling estimator than the SRS estimator and $s_{S\hat{E}}$ is consistently smaller for the stratified sampling variance estimator. For the risk difference, the differences in bias and variance between the stratified sampling and SRS estimators are small, but the stratified sampling variance estimator does have an edge. For the set of simulations run at 750 subjects per domain, the comparisons between the proposed variance estimator for $\hat{\psi}_{MH}$ and the Robins *et al.* variance estimator are mixed. At this sample size, each estimator enjoys an advantage over the other at some of the values of ψ considered. Results in Table 5, though, are uniform. Estimated means squared errors are consistently lower for

the stratified sampling estimator.

In terms of interval estimation, the observed coverage rates are generally not far from the nominal 95% rate for ϕ and δ using either the stratified sampling variance estimator or the SRS estimator. For ψ , observed coverage rates are much further from the nominal rate, especially for the stratified sampling variance estimator and especially as ψ gets large. Based on the bias and variance results, this deviation from nominal coverage can not be blamed on the inaccuracy of the stratified sampling variance estimator. Instead it reflects the low quality of the normal approximation used in interval construction for $\log(\hat{\psi}_{\text{MH}})$ and, to a lesser degree, $\log(\hat{\phi}_{\text{MH}})$.

References

- Birch, M.W. (1964). The detection of partial association, I: the 2×2 case. *Journal of the Royal Statistical Society, B* **26**, 313–324.
- Birch, M.W. (1965). The detection of partial association, II: the general case. *Journal of the Royal Statistical Society, B* **27**, 111–124.
- Breslow, N.E. (1981). Odds ratio estimators when data are sparse. *Biometrika* **68**, 73–84.
- Breslow, N.E., and Day, N.E. (1980). *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.
- Breslow, N.E. and Liang, K.Y. (1982). The variance of the Mantel-Haenszel estimator. *Biometrics* *38*, 943–952.
- Cochran, W.G. (1954). Some methods for strengthening the common χ^2 test. *Biometrics* **10**, 417–451.
- Connett, J., Ejigou, A., McHugh, R., and Breslow, N. (1982). The precision of the Mantel-Haenszel estimator in case-control studies with multiple matching. *American Journal of Epidemiology* **116**, 875–877.
- Donald, A., and Donner, A. (1987). Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine* **6**, 491–499.

- Flanders, W.D. (1985). A new variance estimator for the Mantel-Haenszel odds ratio. *Biometrics* **41**, 637–642.
- Fleiss, J.L. (1984). The Mantel-Haenszel estimator in case-control studies with varying number of controls matched to each case. *American Journal of Epidemiology* **120**, 943–952.
- Gilbaud, O. (1983). On the large-sample distribution of the Mantel-Haenszel odds-ratio estimator. *Biometrics* **39**, 523–525.
- Graubard, B.I., Fears, T.R., and Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control studies. *Biometrics* **45**, 1053–1071.
- Greenland, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Statistics in Medicine* **1**, 217–227.
- Greenland, S. (1989). Generalized Mantel-Haenszel estimators for $K \times J$ tables. *Biometrics* **45**, 183–191.
- Greenland, S, and Robins, J. (1985) Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **41**, 55–68.
- Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489–504.
- Hauck, W.W. (1979). The large-sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* **35**, 817–819.
- Hopkins, C.E. and Gross, A.J. (1971). A generalization of Cochran's procedure for the combining of $r \times c$ contingency tables. *Statistica Neerlandica* **25**, 57–62.
- The Iowa Persian Gulf Study Group (1997). Self-reported illness and health status among Gulf War veterans: a population-based study. *Journal of the American Medical Association* **277**, 238–245.
- Jones, M.F., Doebbeling, B.N, Hall, D.B., Snyders, T.L., Barrett, D.H., Williams, A., Falter, K.H., Torner, J.C., Burmeister, L.F., Woolson, R.F., Merchant, J.A., and Schwartz, D.A. (1998). Methodologic issues in a Population-based Health Survey of Gulf War Veterans. University of Iowa Department of Preventive Medicine and Environmental Health Technical Report No. 98-1.
- Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont, CA.
- Kuritz, S.J., Landis, J.R., and Koch, G.G. (1988). A general overview of Mantel-Haenszel methods: applications and recent developments. *Annual Review of Public Health* **9**, 123–160.

- Landis, J.R., Heyman, E.R., and Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *International Statistical Review* **46**, 237–254.
- Liang, K.Y. (1987). Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics* **43**, 289–299.
- Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* **58**, 690–700.
- Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.
- Mickey, R.M., and Elashoff, R.M. (1985). A generalization of the Mantel-Haenszel estimator of partial association for $2 \times J \times K$ tables. *Biometrics* **41**, 623–635.
- Nurminen, M. (1981). Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika* **68**, 525–530.
- O’Gorman, T.W., Woolson, R.F., and Jones, M.P. (1994). A comparison of two methods of estimating a common risk difference in a stratified analysis of a multicenter clinical trial. *Controlled Clinical Trials* **15**, 135–153.
- Phillips, A., and Holland, P.W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics* **43**, 425–431.
- Robins, J., Breslow, N., and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**, 311–323.
- Rothman, K.J., and Boice, J.D. (1979). *Epidemiologic Analysis with a Programmable Calculator*. U.S. Government Printing Office, Washington, DC.
- Sato, T. (1990). Confidence intervals for effect parameters common in cancer epidemiology. *Environmental Health Perspectives* **87**, 95–101.
- Shah, B.V., Folsom, R.E., LaVange, L.M., Wheelless, S.C., Boyle, K.E., and Williams, R.L. (1995). Statistical methods and mathematical algorithms used in SUDAAN. Research Triangle Institute.
- Sen, P.K. (1988). Combination of statistical tests for multivariate hypotheses against restricted alternatives. In *Advances in Multivariate Statistical Analysis* (eds. S. Dasgupta and J.K. Ghosh), Indian Statistical Institute, Calcutta, 377–402.
- Sugiura, N. and Otake, M. (1974). An extension of the Mantel-Haenszel procedure to $K \times 2 \times C$ contingency tables and the relation to the logit model. *Communi-*

cations in Statistics **3**, 829–842.

Tarone, R.E. (191). On summary estimators of relative risk. *Journal of Chronic Disease* **34**, 463–468.

Ury, H. K. (1982). Hauck's approximate large-sample variance of the Mantel-Haenszel estimator. *Biometrics* **38**, 1094–1095.

Weerasekera, D.R., and Bennett, S. (1992). Adjustments to the Mantel-Haenszel test for data from stratified multistage surveys. *Statistics in Medicine* **11**, 603–616.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **66**, 411–414.

Yanagawa, T. and Fujii, Y. (1990). Homogeneity test with a generalized Mantel-Haenszel estimator in $L \times J$ contingency tables. *Journal of the American Statistical Association* **85**, 744–748.

Yanagawa, T. and Fujii, Y. (1995). Projection-method Mantel-Haenszel estimator for $K \times J$ tables. *Journal of the American Statistical Association* **90**, 649–656.