Ann. Rev. Public Health. 1988. 9:123–60

A GENERAL OVERVIEW OF MANTEL-HAENSZEL METHODS: Applications and Recent Developments

Stephen J. Kuritz

Hoechst/Celanese Specialties Group, 1 Main Street, Chatham, New Jersey 07928

J. Richard Landis

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109

Gary G. Koch

Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27514

1. INTRODUCTION

Many health research investigations are concerned with the relationship between a primary factor, such as a potentially harmful exposure, a new therapy, or an intervention, and a response variable such as disease status, level of functioning, or extent of improvement. Quite frequently, both of these variables are reported on categorical measurement scales, either nominal or ordinal, so that the resulting data can be displayed as observed frequencies in a two-way contingency table. However, this factor-response relationship also may be influenced by additional variables, typically referred to as intervening variables, confounding variables, effect modifiers, or covariables, because of their effects on the response variable and its relationship with the primary factor. For example, the study may be conducted at several institutions involving different investigators and patient populations;

0163-7525/88/0510-0123$02.00

also, other risk factors may provide important sources of variation in the data, or one or more socio-demographic variable such as age, race, sex, or family income may be highly associated with the response variable. Definitive assessment of the relationship between a primary risk factor and the response variable requires that appropriate adjustments for these additional sources of variation be utilized either in the design of such studies and/or in the analysis of the resulting data.

Historically, it was the need to combine information relating a primary factor and a response variable across a set of (2×2) contingency tables based on the cross-classification of other relevant variables that led to the development of several related procedures, now popularly referred to as "Mantel-Haenszel methods." In a classic paper on "methods to strengthen the common chi-square statistic," Cochran (11) proposed a test involving a weighted mean difference across the set of (2×2) tables. His derivation was linked to binomial model assumptions, which resulted in the requirement of a moderately large sample size N_h for each table. From another point of view, Mantel & Haenszel (43), in a landmark paper concerning retrospective studies, noted that this same problem could be approached using a hypergeometric probability model, which permits either exact tests or requires only the overall sample size to be reasonably large for asymptotic results to be applicable. Since the resulting test statistics from these two procedures differ only by a term of $(N_h - 1)/N_h$ within the hth table and a continuity correction overall, they are nearly identical when the sample sizes are moderate to large in each table ($N_h \geq 20$). Moreover, the Mantel-Haenszel test statistic has the desirable property that it is entirely appropriate even with sample sizes as small as 2, as in matched pair studies, provided that there are enough tables. Birch (3) demonstrated that under the assumption of a constant odds ratio within each of the respective tables, the Mantel-Haenszel test statistic is the uniformly most powerful unbiased test. Also, it is asymptotically equivalent to likelihood ratio tests from unconditional logistic regression with effects for stratum and the factor when sample sizes are large for each table, and from conditional logistic regression when they are small (7).

The primary purpose of using the Mantel-Haenszel procedure is to provide a test with statistical power directed at alternatives encompassing an average effect of the factor on the response across strata based on the set of covariates of interest. In many situations, the sample sizes for some tables may be sparse, the magnitude of the association between factor and response effect may vary across tables, and the association may be small within a table due to the stratification by other important covariates. Nevertheless, if the association is slight but consistent across the tables, this procedure will be effective in detecting that association.

Perhaps the most important distinguishing feature of the Mantel-Haenszel

(43) test statistic is its connection to randomization model considerations. Quite frequently, the data for health research investigations are not obtained through well-controlled, probability sampling from a target population. More typically, these data are collected under a variety of observational designs such as case-control studies, with or without matching or convenience sampling for a randomized, multicenter efficacy trial. Since assumptions of random sampling are not tenable, analytical strategies based on nonparametric randomization methods are more appropriate than postulated model-based methods assuming a known sampling link to a larger reference population. For such situations the Mantel-Haenszel testing procedure does provide such a randomization, design-based approach to hypothesis testing. These methods require no assumptions other than the randomization of subjects to levels of the factor, either explicitly as in randomized controlled clinical trials, or implicitly by hypothesis or from conditional distribution arguments for observational data from restrictive populations such as retrospective studies, prospective nonrandomized studies or case-control studies (32, 35).

An issue for interpretation of Mantel-Haenszel methods is their possibly limited scope for inference, since in a strict statistical sense the conclusions might apply only to the study sample; generalizations to a target population by them or any other method would require nonstatistical arguments concerning the representativeness of the study subjects to the cross-section of individuals in the target population. These issues of "extended inference," in contrast to "local inference," are discussed in more detail in Koch et al (31) and Koch & Gillings (32).

In this paper we review the Mantel-Haenszel methods, both for hypothesis testing and estimation of an average odds ratio, for a set of (2×2) tables. Extensive details on the recently published variance formulae for this average odds ratio are provided using a unified set of notation (6, 8, 12, 17, 19, 26, 28, 48, 49, 54). The applications of these methods for investigating treatment differences and change in response over time within treatment subgroups are illustrated using data from a randomized multicenter clinical trial. More generally, the extensions of this methodology to factor-response situations and repeated measurement designs involving a set of $(s \times r)$ contingency tables are summarized. In order to de-emphasize the necessary mathematical details, particularly the matrix formulations of these methods, we have outlined these developments in Appendix 1. These generalized Mantel-Haenszel methods are illustrated within the context of the same data set, comparing three treatments relative to a summary measure of response over time and investigating the change in the response profile over time within treatment subgroups.

Whereas these Mantel-Haenszel methods, both for hypothesis testing and estimation, address the notion of an "average effect" across strata, it is

desirable to assess whether the individual table effects are homogeneous across strata or whether the relationship between factor and response varies greatly across the tables. For a set of (2×2) tables, several methods are available to test for the presence of such interaction in terms of variation of stratum-specific odds ratios; some strategies for this purpose are reviewed in Gart (24). More recently, a family of test statistics for assessing homogeneity of the odds ratios across strata under minimal assumptions and with straight forward computing is described in Breslow & Day (7). In settings where sample sizes are adequate within each table, additive log-linear models can be fit to the data and interaction evaluated through tests for their goodness of fit (5). In the context of the Mantel-Haenszel methods, Koch et al (33) describe a pseudo-homogeneity test statistic for this purpose that requires both adequate sample sizes within strata and caution with regard to its interpretation. Otherwise, an exact procedure based on the extended hypergeometric distribution is described in Gart (24) and has been implemented in a computer program by Thomas (53a). These issues relating to interaction also are discussed in Fleiss (18), Gart (24), Kleinbaum et al (30), and Schlesselman (51).

This review paper necessarily is incomplete, when considering all the applications and developments of the Mantel-Haenszel methods that have been presented in the literature over the years since the original paper appeared in 1959. Rather than attempting to be exhaustive, we have selected a complex data set involving multiple treatment groups, baseline covariates, and an ordinal response variable measured repeatedly over time on the same subjects, as a mechanism to illustrate the flexibility of these methods to investigate a wide variety of hypotheses.

2. NOTATION AND GENERAL METHODOLOGY

Suppose the subjects are classified according to a primary factor such as (*a*) treatments in either an unmatched or repeated measures design, (*b*) case/control status in a matched design, or (*c*) levels of time in a longitudinal study. Suppose further that the relationship between this primary factor and the response variable of interest may be influenced by a set of intervening variables that are considered to be potential sources of bias. For example, in unmatched study designs these covariates may include centers in a multicenter clinical trial or pretreatment severity; whereas in repeated measures designs these covariate levels are either the individual subject or the matched set. Let the strata be indexed by $h = 1, 2, \ldots, t$ and let n_{hij} denote the number of subjects in the sample who are jointly classified as belonging to the ith factor level, the jth level of the response variable, and the hth stratum. The resulting ($s \times r$) contingency table for the hth stratum is summarized in Table 1.

Table 1 Observed contingency table for stratum h

Factor levels	Response variable categories 1	2	· · ·	r	Total
1	n_{h11}	n_{h12}	· · ·	n_{h1r}	$N_{h1\cdot}$
1	n_{h21}	n_{h22}	· · ·	n_{h2r}	$N_{h2\cdot}$
·	·	·	· · ·	·	·
s	n_{hs1}	n_{hs2}	· · ·	n_{hsr}	$N_{hs\cdot}$
Total	$N_{h\cdot1}$	$N_{h\cdot2}$	· · ·	$N_{h\cdot r}$	N_h

For hypotheses involving the association between a row factor such as treatment and a column response variable, $h = 1, 2, \ldots, t$ indexes the stratification determined by the cross-classification of the relevant covariables. In contrast, for a repeated measurement design, a primary hypothesis involves homogeneity of the marginal distribution of the response variable across the levels of the repeated measurement dimension. For example, suppose that vision grade of the left and right eyes is to be compared in a sample of individuals for whom each eye is classified as good, moderate, and poor. In this context, $i = 1, 2$ indexes the left and right eye, $j = 1, 2, 3$ indexes the level of vision, and $h = 1, 2, \ldots, t$ indexes the individuals. Or, suppose that a retrospective case/control study matched each diseased case to both a hospital and a neighborhood control. Then, the primary hypothesis is that the distribution of exposure to the risk factor of interest is the same for the cases and matched controls. Consequently, in this situation $i = 1, 2, 3$ indexes the repeated measures factor of case, hospital control, and neighborhood control for the respective matched sets, j indexes the levels of the risk factor, and the strata are the unique matched sets indexed by $h = 1, 2, \ldots, t$.

If we assume that the row marginal totals $\{N_{hi\cdot}\}$ and the column marginal totals $\{N_{h\cdot j}\}$ in Table 1 are fixed either by sampling design or conditional distribution arguments, the overall null hypothesis of "no partial association" between the row dimension (factor) and the column dimension (response) can be stated as

H_0: For each of the stratum levels indexed by $h = 1, 2, \ldots, t$, the response variable is distributed at random with respect to the factor levels.

In other words, from a finite population sampling perspective, this null

hypothesis posits that the data in the respective rows of the hth stratum can be regarded as a successive set of simple random samples of sizes $\{N_{hi\cdot}\}$ from a fixed population corresponding to the marginal total distribution of the response variable $\{N_{hj}\}$.

Under this null hypothesis, it can be shown that the observed frequencies, $\{n_{hij}\}$, follow the probability model,

$$Pr(\{n_{hij}\}|H_0) = \frac{\prod_{i=1}^{s} N_{hi\cdot}! \prod_{j=1}^{r} N_{h\cdot j}!}{N_h! \prod_{i=1}^{s} \prod_{j=1}^{r} n_{hij}!}. \qquad 1.$$

Expression 1 simplifies to the familiar probability used in Fisher's exact test for a single (2 × 2) table. Under this hypothesis of randomness, we can compute the expected value for each frequency and the covariance of each frequency with each of the other frequencies in Table 1 as outlined in Appendix 1. Using these quantities, we can investigate a series of methods directed at alternative hypotheses involving "average effects" of the primary factor on the distribution of the response variable, adjusted for the potential stratum effects. These alternative hypotheses, and the corresponding test statistics, in terms of large-sample quadratic form test statistics, are presented below in a sequence reflecting additional refinement of the hypothesis by incorporating scores for ordinal levels of the response variable and/or the row factor. Other alternative hypotheses, such as those directed at multivariate functions of the response profile, are not considered here, although they have been described elsewhere; e.g. Koch et al (31–33).

2.1 Alternative Hypothesis: General Association

In the most general case, we are interested in the extent to which H_0 can be rejected in favor of the distribution of the response variable differing in nonspecific patterns across levels of the row factor, adjusted for the covariates. Here the levels of both the response variable and the factor are incorporated into the analysis as nominal scale variables. As noted in Appendix 1, the generalized Mantel-Haenszel test statistic, with d.f.$=(r-1)(s-1)$ $Q_{MH(1)}$ in (A1.6), is based on the sums over the t tables of the differences between the observed and expected frequencies relative to the sum of the covariance matrices.

In unmatched studies, for the special case in which $s = r = 2$, the resulting data can be summarized in a set of t: (2 × 2) tables. Here $Q_{MH(1)}$ is identical to the test statistic proposed in Birch (3), is identical (except for the lack of a continuity correction) to the statistic recommended in Mantel & Haenszel

(43), and differs from the test statistic proposed in Cochran (11) only by a factor of $[(N_h - 1)/N_h]$ in the variance term for each table. In repeated measurement or matched designs, $Q_{MH(1)}$ simplifies to a number of familiar test statistics as noted in White et al (55) and Somes (52). In particular, if the response variable is dichotomous ($r = 2$), $Q_{MH(1)}$ is equivalent to McNemar's (45) test when $s = 2$ and to Cochran's (10) Q criterion when $s > 2$. Furthermore, for $r > 2$ and $s > 2$, this result is identical to the "Lagrange multiplier" test derived in Birch (4), the test of interchangeability due to Madansky (38), and the extended Mantel-Haenszel criterion described in Darroch (13), Mantel & Byar (41), and White et al (55).

2.2. *Alternative Hypothesis: Mean Responses Differ*

In situations involving an ordinal response variable, we are interested in the extent to which measures of location, reflected often as average response, differ across the factor levels. The average response for each level of the factor can be generated by assigning column scores, say $a_{h1}, a_{h2}, \ldots, a_{hr}$, and forming the mean score,

$$F_{hi} = \sum_{j=1}^{r} a_{hj}[n_{hij}/N_{hi\cdot}],$$

for the ith row. The specific choice of the scores are not discussed here; further details are available in Landis et al (36), Koch et al (33), and Koch & Edwards (34). As summarized in Appendix 1, the differences among these mean scores across the subtables, relative to their covariances, can be used to create a test statistic directed at differences in the average response among the factor levels. In this context, $Q_{MH(2)}$ in (A1.10) reflects the extent to which the mean scores for certain levels of the factor consistently exceed (or are exceeded by) the mean scores for other levels of the factor, and thus is directed at the average of such differences across the t strata. In particular, for $s = 2$, this test is identical to the extended Mantel-Haenszel test statistic proposed in Mantel (39). Moreover, if marginal rank or ridit-type scores are obtained from each table with midranks assigned for ties, $Q_{MH(2)}$ is equivalent to an extention of the Kruskal-Wallis analysis of variance (ANOVA) test on ranks, conditioning on the levels of the strata; for $s = 2$, this is the van Eltern (54a) test, for which additional discussion is given in Lehmann (36a).

With respect to repeated measures designs, this test for differences among the mean scores is equivalent to the Mantel-Haenszel statistic provided in Breslow & Day (7) for matched case-control data with an ordinal risk factor. Where marginal rank scores are assigned within each stratum of a repeated measures design, $Q_{MH(2)}$ simplifies to the chi-square criterion proposed by Friedman (20) from a two-way rank ANOVA within blocks for subjects. In

both designs, $Q_{MH(2)}$ with degrees of freedom (d.f.) = $(s - 1)$ provides increased statistical power to detect departures from homogeneity across factor levels for an ordered response variable relative to $Q_{MH(1)}$ with d.f. = $(s - 1)(r - 1)$, although more complex patterns of association will not necessarily be detected.

2.3. *Alternative Hypothesis: Linear Trend in Mean Responses*

For many epidemiological and clinical studies, not only is the response variable frequently measured on an ordinal scale, but the factor levels also are ordered. Suppose we assign row scores, say c_{h1}, c_{h2}, ..., c_{hs}, to the ordinal levels of the factor. Then, as outlined in Appendix 1, we can develop a generalized Mantel-Haenszel test statistic with d.f. = 1 directed at the extent to which there is a consistent positive (or negative) association between the response scores and the factor level scores in the respective strata. Specifically, $Q_{MH(3)}$ in (A1.14) is directed at the extent to which H_0 is contradicted in favor of a linear progression in the average response across the levels of the factor relative to the assigned scores.

This statistic, $Q_{MH(3)}$, is identical to the correlation statistic proposed by Mantel (39) and Birch (4). If marginal rank or ridit-type scores are assigned to both the rows and columns of each table with midranks assigned for ties, this statistic is equivalent to an extension of the Spearman rank correlation test, conditioning on the levels of the covariates. With only 1 degree of freedom, this statistic has increased power relative to either $Q_{MH(1)}$ or $Q_{MH(2)}$ for linear correlation alternatives relative to the assigned scores.

3. FACTOR-RESPONSE HYPOTHESES

Quite often the primary hypothesis concerns the relationship between a factor and a response variable in the presence of a number of covariates. If the factor has s levels and the response has r levels, the data at each stratum of the covariate set can be summarized in an $(s \times r)$ contingency table as in Table 1. Each subject is classified into exactly one cell of this three-way table, corresponding to the hth level of the covariate set and the (i,j)th level of the factor-response structure. These methods will be described in more detail, first beginning with the familiar (2×2) table layout.

3.1. *(2 × 2) Tables*

In the simplest case, both the factor and the response are dichotomous and the frequency data can be arranged in a series of t (2×2) tables. In this setting all three test statistics outlined in Appendix 1 are identical, having d.f. = 1. As noted above, the test statistic in Eq. A1.6 from Appendix 1 simplifies (except for the lack of a continuity correction) to the familiar Mantel-Haenszel statistic,

$$Q_{\mathrm{MH}(1)} = \frac{\left[\sum_{h=1}^{t} n_{h11} - \sum_{h=1}^{t} \frac{N_{h1\cdot}N_{h\cdot 1}}{N_h}\right]^2}{\sum_{h=1}^{t} \frac{N_{h1\cdot}N_{h2\cdot}N_{h\cdot 1}N_{h\cdot 2}}{N_h{}^2(N_h - 1)}}. \qquad 2.$$

Guidelines concerning sample size requirements in order for the chi-square approximation for $Q_{\mathrm{MH}(1)}$ to be appropriate are provided by Mantel & Fleiss (42); briefly, the sum across the t strata of the observed frequency for each cell of the (2 × 2) table should have an expected value exceeding 5 and an allowable range of 5 on each side of the expected value. Further sample size discussions are available in Breslow & Day (7), Fleiss (18), and Koch et al (33).

The statistic in Expression 2 can be viewed as a test directed at the alternative hypothesis that a weighted average of the stratum-specific odds ratios, say $\bar{\psi} = \Sigma_h w_h \psi_h / \Sigma_h w_h$, differs from 1, the expected value under H_0. Having rejected H_0, the obvious next step is to estimate the magnitude of this average factor-response association. The most widely accepted estimator for $\bar{\psi}$,

$$\hat{\bar{\psi}}_{\mathrm{MH}} = \frac{\sum_{h=1}^{t} n_{h11}n_{h22}/N_h}{\sum_{h=1}^{t} n_{h21}n_{h12}/N_h}, \qquad 3.$$

was proposed in the original Mantel-Haenszel paper (43). In addition, there have been many estimators proposed for the common odds ratio, ψ, under the assumption of homogeneity of the stratum-specific odds ratios. Among these alternative estimators are the unconditional maximum likelihood (21, 24), Woolf (56), and the Modified Woolf (22, 25) estimator. These estimators require large numbers of observations in each stratum, so that asymptotic properties arise as the number of strata remains fixed while the total number of subjects increases without bound, corresponding to standard large sample theory. In contrast, the conditional maximum likelihood estimator (3, 23) is appropriate with as few as two observations per stratum (e.g. matched pair studies) so long as the number of strata is sufficiently large.

All of these estimators are consistent, asymptotically normal, and (at $\psi = 1$ for $\hat{\psi}_{\mathrm{MH}}$) asymptotically efficient (6, 53). Simulation work by McKinlay (44) and Hauck et al (29) demonstrated that there was little difference among the estimators in terms of bias and precision for the unmatched design with large numbers of subjects in each stratum. The Woolf and Modified Woolf statistics

were shown to have a large bias problem as stratum sample sizes became more moderate. When considering ease of computation together with statistical properties, McKinlay (44) recommended using the Mantel-Haenszel estimator in Expression 3.

All of these estimators, except for the Mantel-Haenszel statistic in Expression 3, require the assumption of a common odds ratio for each stratum. The Mantel-Haenszel estimate was intended to be a weighted average of the individual odds ratios (40), where the weights are related to the precision of $\hat{\psi}_h$ (7, 15). In fact, Mantel & Haenszel (43) indicated their disbelief in the constancy of the underlying odds ratio, stating that the assumption of a constant relative risk is usually untenable. McKinlay (44) interpreted an adjusted odds ratio as in Expression 3 estimated from a stratified analysis as representing the constant component of an association across strata; whereas Landis et al (35) and others refer to this quantity as "average partial association." Even though the stratum-specific odds ratios may be heterogeneous, one often is interested in a summary measure (27). However, most investigators generally agree that a combined odds ratio, which is based on individual odds ratios that differ substantially in direction, some being less than unity and others greater than unity, can be difficult to interpret and perhaps should not be used.

Formal statistical tests for this hypothesis of homogeneous odds ratios across strata can be conducted within the context of fitting an additive logit-linear model with either maximum likelihood or weighted least squares methods, provided the within-stratum sample sizes are adequate. Details of these methods are provided in Koch et al (33) and are not discussed further in this review. Under an extension of the hypergeometric probability model in Expression 1, it is possible to develop a test statistic for this homogeneous odds ratio hypothesis that does not require logit-linear model fitting. In particular, Breslow & Day (7) proposed a test statistic directed at the potential lack of homogeneity of odds ratios,

$$Q_{\mathrm{BD}} = \sum_{h=1}^{t} \frac{[n_{\mathrm{h}11} - E(n_{h11}|\hat{\tilde{\psi}}_{\mathrm{MH}})]^2}{var(n_{h11}|\hat{\tilde{\psi}}_{\mathrm{MH}})}, \qquad 4.$$

where $E(n_{h11}|\hat{\tilde{\psi}}_{\mathrm{MH}})$ is the expected value of $n_{\mathrm{h}11}$ under the hypothesis of homogeneity of odds ratios, $var(n_{h11}|\hat{\tilde{\psi}}_{\mathrm{MH}})$ is the variance of n_{h11} under the same hypothesis, and $\hat{\tilde{\psi}}_{\mathrm{MH}}$ is the estimate of the average odds ratio under this homogeneity hypothesis. Assuming acceptably large sample sizes in each subtable, Q_{BD} follows an approximate chi-square distribution with d.f. $= t - 1$. This test statistic and corresponding p-value may be obtained as part of the output from the FREQ procedure in SAS (50).

For 20 years after the Mantel-Haenszel odds ratio in Expression 3 was first

proposed in 1959, no variance formula was available. More recently, a number of such variance formulae have been proposed. These variance formulae can be classified into one of three categories reflecting the limiting models for which they are consistent asymptotically:

1. Requires number of observations in each stratum (N_h) to be large.
2. Requires total number of strata (t) and observations (N) to be large.
3. Appropriate in either limiting model 1 or 2.

The precise formulations for these variance estimators are presented in Appendix 2. Phillips & Holland (48) note that these variance formulae have been derived by conditioning on the finiteness of the Mantel-Haenszel estimator
ln ($\hat{\psi}_{MH}$) is infinite.

Initially, Hauck (28) proposed a variance formula for $\hat{\psi}_{MH}$ arising from the product binomial model, denoted here by v_H, under the restriction of a common odds ratio and appropriate for the first limiting model (1). Gilbaud (26) provided a variance formula for the less restricted setting (heterogeneity of odds ratios), denoted here by v_{G1}, and also proposed a variance formula (denoted here by v_{G2}) to use when conditioning on all of the margins of each table. Consequently, v_{G1} is linked to product binomial sampling assumptions and v_{G2} is linked to hypergeometric probability distribution assumptions, and both are appropriate only for the first asymptotic limiting model (1). When homogeneity is present (i.e. a common odds ratio across strata), both v_{G1} and v_{G2} simplify to v_H.

Ury (54) noted that v_H would give different results if the rows were reversed in each table, and proposed several methods for symmetrization (such as using the geometric mean), resulting in a symmetrized version of the variance due to Hauck, denoted here by v_{HS}. When there is only one stratum ($t = 1$), all of these formulae simplify to the usual large-sample variance for the odds ratio from a single 2×2 table as presented in Fleiss (18), which is

$$\text{var}(\hat{\psi}) = \hat{\psi}^2\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

Breslow (6) proposed a variance formula for the restricted setting of homogeneous odds ratios that is linked to hypergeometric probability distribution assumptions. This variance, denoted here by v_{B1}, is appropriate for both asymptotic limiting models (1 and 2). However, Breslow noted that the calculations required for v_{B1} are "... rather laborious," and proposed an empirical version for this variance formula (denoted here by v_{B2}) that is consistent only under the second asymptotic limiting model (2).

For the matched pairs design, in which the Mantel-Haenszel odds ratio is identical to the conditional maximum likelihood estimator, v_{B2} simplifies to the well-known variance formula involving only the discordant pairs (16, 18). Connett et al (12) presented an explicit, relatively simple expression for v_{B2} derived from a case-control study in which each case was matched to the same number of controls, and Fleiss (19) generalized their results to studies in which there are varying numbers of controls per case. Phillips & Holland (48) also derived a variance estimator valid only for this setting.

A number of variance formulae for the Mantel-Haenszel odds ratio that are appropriate in both limiting models have been proposed recently. As noted in the preceding section, v_{B1} is consistent in both limiting models but, based on hypergeometric variances, requires tedious computations. Breslow & Liang (8) suggested forming a generalized variance estimator by computing a weighted average of v_H and v_{B2}, where the ad-hoc weights are N_h for v_H and t^2 for v_{B2}. A modification of this combined variance formula, denoted here by v_{CS}, substitutes v_{HS} for v_H in computing the weighted average. Breslow & Liang (8) recommend using v_{CS} rather than v_C.

Flanders (17) proposed a variance formula (denoted here by v_F) which is essentially a modification of v_H in order to be consistent in both limiting models. A different estimator that also is consistent in both limiting models was proposed by Robins et al (49), and is denoted here by v_R. These two variance formulae are very similar, and both assume a common odds ratio. In fact, for matched sets with one case and a constant number of controls in each stratum, v_F is identical to v_R. Robins et al (49) also proposed a symmetrized version of v_R (denoted here by v_{RS}), where v_{RS} is the arithmetic average of v_R computed first on the original tables, and then recomputed after interchanging the rows of each table.

Ideally, one would like to select one of these variance formula as "best" under general criteria. Simulation studies have been conducted for some of these formulae (8, 17, 49). In general, these results indicate that the variances do well within the limiting models for which they are appropriate, and that v_{HS} and v_{CS} perform better than their nonsymmetric counterparts v_H and v_C. The formulae proposed by Flanders and by Robins et al fared well in all settings, and since they are appropriate in both limiting models they appear to be the formulae of choice for computing the variance of the Mantel-Haenszel odds ratio. Various characteristics of each of these variance formulae are summarized in Table 2.

3.2. Example

The data in Table 3 were obtained from a placebo-controlled, randomized, multicenter clinical trial comparing the effects of a combination drug labeled (A & B) with its separate components A and B on the level of obstetrical

Table 2 Characteristics of variance formulae for M-H odds ratio

Variance formula	Underlying model	Assumes common odds ratio	Symmetric	Asymptotic consistency[a]
v_H	Product binomial	Yes	No	I
v_{HS}	Product binomial	Yes	Yes	I
v_{G1}	Product binomial	No	No	I
v_{G2}	Hypergeometric	No	No	I
v_{B1}	Hypergeometric	Yes	No	III
v_{B2}	Hypergeometric	Yes	Yes	II
v_C	Mixed	Yes	No	III
v_{CS}	Mixed	Yes	Yes	III
v_F	Hypergeometric	Yes	No	III
v_R	Product binomial	Yes	No	III
v_{RS}	Product binomial	Yes	Yes	III

[a]I. Requires number of observations in each stratum (N_h) to be large; II. Requires total number of strata (t) and observations (N) to be large; III. Consistent in either limiting model I or II.

related pain in women who recently had delivered a baby. For our purposes in this review, we have limited attention to three treatment subgroups: (A & B), A Only, and Placebo. The randomly assigned treatments were administered immediately after delivery and again at four hours after delivery.

The response variable for this study consisted of self-reported levels of pain measured on a five point ordinal scale with categories O:None, 1:Little, 2:Some, 3:Lots, and 4:Terrible. These measurements were recorded immediately after delivery and at each hour for 8 hr following delivery. The initial pain status immediately after delivery, but prior to treatment, was reported as either 2:Some or 3:Lots, despite the potential levels of 0–4. The eight strata correspond to the cross-classification of initial pain status by each of four different medical centers. These data are described further in Koch et al (33).

To illustrate the Mantel-Haenszel methods for (2 × 2) tables, first we consider a factor and a response variable each at two levels as displayed in Table 3. The factor levels correspond to either the combination drug (A & B) or the Placebo treatment and the response variable was created to reflect a combination both of the intensity and duration of post-partum pain. Each patient was classified as having either 0–4 or 5–8 hr of at least Some pain (levels 2–4). The resulting (2 × 2) table is summarized for each of the eight strata. Over all the eight strata, 36% of women receiving placebo treatment reported 5–8 hr with at least Some pain compared to only 8% of women receiving drugs A & B, for an (unadjusted) odds ratio of 6.3.

Note that although the odds ratios from the eight strata indicate that drugs A & B constitute a more effective treatment for minimizing this measure of post-partum pain than the placebo treatment, they differ in magnitude from a

Table 3 Distribution of total number of hours of at least some port-partum pain[a] by treatments and strata: Dichotomous factor and response

Stratum number	Covariates: Center	Covariates: Initial pain[b]	Treatment	Hours with at least Some post-partum pain: 0–4	Hours with at least Some post-partum pain: 5–8	Total	Prop. 5–8 hours	Est. odds ratio
1	1	Some	A & B	19	0	19	0.00	12.1
			Placebo	14	4	18	0.22	
2	1	Lots	A & B	25	1	26	0.04	18.3
			Placebo	15	11	26	0.42	
3	2	Some	A & B	19	1	20	0.05	6.3
			Placebo	15	5	20	0.25	
4	2	Lots	A & B	25	1	26	0.04	29.2
			Placebo	12	14	26	0.54	
5	3	Some	A & B	24	1	25	0.04	8.5
			Placebo	17	6	23	0.26	
6	3	Lots	A & B	20	6	26	0.23	2.8
			Placebo	12	10	22	0.45	
7	4	Some	A & B	20	1	21	0.05	3.8
			Placebo	16	3	19	0.16	
8	4	Lots	A & B	18	4	22	0.18	4.5
			Placebo	8	8	16	0.50	
None			A & B	170	15	185	0.08	6.3
			Placebo	109	61	170	0.36	

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; at least Some pain obtained by combining levels 2–4.
[b]Initial pain status reported only as either 2:Some or 3:Lots, despite potential levels of 0–4.

low of 2.8 (stratum 6) to a high of 29.2 (stratum 4). Using the Breslow-Day test for homogeneity of odds ratios in Expression 4, we obtain $Q_{BD} = 6.45$ with d.f. = 7 ($p = 0.49$). Thus, despite this wide range in the magnitude of the stratum-specific odds ratios, these data do not provide sufficient evidence to reject the hypothesis of homogenous odds ratios.

The Mantel-Haenszel (M-H) average odds ratio in Expression 3, adjusting for these strata effects, is $\hat{\psi}_{MH} = 7.2$, somewhat larger than the unadjusted estimate of 6.3. Estimated standard errors for the M-H average odds ratio on the natural logarithm scale are presented in Table 4, together with 95%

Table 4 Estimated standard errors for ln Mantel-Haenszel (MH) average odds ratio and 95% confidence intervals for MH average odds ratio for the post-partem pain data in a factor-response setting

Variance formula	Estimated standard error for ln MH avg. odds ratio	95% Confidence interval for MH avg. odds ratio
G1	0.32	(3.89, 13.48)
G2	0.30	(3.99, 13.15)
HS	0.34	(3.71, 14.13)
B2	0.33	(3.77, 13.91)
CS	0.34	(3.72, 14.09)
F	0.33	(3.82, 13.75)
R	0.34	(3.74, 14.04)
	Alternative methods	
Logit	0.34	(3.12, 11.83)
Test-based	0.30	(4.00, 13.11)
Extended hypergeometric:		
Approximate		(3.71, 14.66)
Exact		(3.63, 14.26)

confidence intervals for the M-H average odds ratio using the selected variance formulae from Appendix 2. For these data, the standard errors and confidence intervals are all very similar, primarily due to the modest sample size in each row of each (2 × 2) table.

These confidence intervals may be compared to those obtained from methods that are more well known. For example, 95% confidence intervals for the odds ratio obtained from four alternative methods are presented in the lower section of Table 4. For example, the logit estimates of the common odds ratio can be obtained directly by fitting a logistic regression model with parameters for treatment and strata. The resulting 95% confidence limits are routinely provided in the output generated by the FREQ procedure in SAS. The smaller values for these limits reflect the fact that the logit odds ratio estimate is 6.1, compared to 7.2 for the Mantel-Haenszel estimate. The upper and lower confidence limits for the test-based approach due to Miettinen (47) are now part of the output produced from the FREQ procedure in SAS (50). The caveats concerning the appropriateness of this ad hoc method are described in Kleinbaum et al (30); briefly, it may only be appropriate near the null value of 1. For these data, the test-based confidence limits most closely resemble those obtained from the variance formula proposed by Gilbaud (26), v_{G2}. Gart (24) reviewed procedures (23) based on the extended hypergeometric probability distribution that can be used to compute both approximate and exact confidence limits for a common odds ratio. Thomas (53a) developed

computer programs for calculating these quantities. These results are summarized in the last two lines of Table 4. Each of these intervals is slightly longer than those obtained from the noniterative formulae presented in the upper section of the table.

Returning to the randomization model test for the statistical significance of the treatment effect, we note that $Q_{MH(1)} = 42.80$ with d.f. = 1 ($p < 0.01$). We reject the null hypothesis of "no partial association" in favor of the alternate hypothesis suggesting that the marginal distributions of the response variable (0–4 vs 5–8 hr with at least Some pain) are not homogeneous for both treatment regimens (drugs A & B vs Placebo) on average across the eight strata. The unadjusted test statistic obtained by collapsing the data across clinics and baseline status shown at the bottom of Table 3 is Q = 40.50 with d.f. = 1 ($p<0.01$). Thus, although the odds ratios exhibit considerable variation across the strata, the effect of stratification on the test statistic and the odds ratio estimate is minimal in this instance. In practice, support for using an unadjusted test statistic would require fitting a reduced logit-linear model. Here we are only performing that test for comparative purposes.

3.3 General s × r *Tables*

Quite often, either the factor or the response variable is measured on scales with more than two levels. Suppose we let s be the number of factor levels and r be the number of response levels. Then the resulting frequency data can be displayed in an ($s \times r$) contingency table. For example, consider the data in Table 5 resulting from the same post-partum pain data summarized in Table 3, now including a third treatment regimen, drug A only. In addition, each woman is classified according to the actual number of hours of at least Some pain during the 8 hr of follow-up, rather than the dichotomized duration illustrated in Table 3. Thus, the strata are the same as presented in Table 3, but the data summarized in Table 5 now consist of eight (3×9) contingency tables.

Estimation of treatment effects is much more complex for the general ($s \times r$) table situation, particularly if one or both of the dimensions of the tables are not ordinally scaled. Recently, investigators have been developing some extensions of the Mantel-Haenszel estimating procedure (14, 37, 46), although none appear to be fully generalizable to a series of ($s \times r$) tables. Ordinal models provide summary measures (1, 9), but the sample size requirements and structural model assumptions, such as proportional odds, are much more restrictive than required for the Mantel-Haenszel methods to be applied. For these reasons, we focus in the remainder of this paper on the hypothesis-testing aspects of the Mantel-Haenszel methods, rather than on estimating combined measures of association.

The data in Table 5 permit consideration of all three alternative hypotheses

Table 5 Distribution of total number of hours of at least some port-partum pain[a] by treatments and strata: multi-level factor and response

Stratum number	Covariates: Center	Covariates: Intital pain	Treat-ment	Post-partum pain hours: 0	1	2	3	4	5	6	7	8	Total	Observed mean score
1	1	Some	A & B	6	4	5	3	1	0	0	0	0	19	1.42
			A Only	1	5	4	2	1	2	0	1	2	18	3.11
			Placebo	2	4	4	2	2	0	3	0	1	18	2.89
2	1	Lots	A & B	2	6	9	7	1	1	0	0	0	26	2.08
			A Only	0	1	7	4	4	2	1	3	6	28	4.57
			Placebo	0	3	7	3	2	2	2	1	6	26	4.27
3	2	Some	A & B	5	7	4	0	3	1	0	0	0	20	1.60
			A Only	7	1	8	1	1	1	0	0	0	19	1.53
			Placebo	4	5	2	1	3	1	2	0	2	20	2.85
4	2	Lots	A & B	5	9	8	2	1	0	0	1	0	26	1.62
			A Only	1	7	9	2	3	0	0	1	3	26	2.85
			Placebo	2	2	3	2	3	2	3	2	7	26	4.77
5	3	Some	A & B	13	4	2	4	1	1	0	0	0	25	1.16
			A Only	6	3	3	5	3	1	0	0	1	22	2.23
			Placebo	5	4	4	1	3	1	0	0	5	23	3.13
6	3	Lots	A & B	4	9	5	2	0	0	1	2	3	26	2.65
			A Only	3	2	3	1	1	5	1	2	4	22	4.18
			Placebo	1	2	1	6	2	2	2	0	6	22	4.55
7	4	Some	A & B	14	2	2	2	0	0	0	0	1	21	0.95
			A Only	13	1	2	2	0	1	0	0	0	19	0.84
			Placebo	10	0	1	1	4	1	1	0	1	19	2.11
8	4	Lots	A & B	8	4	3	2	1	0	3	0	1	22	2.09
			A Only	6	3	1	6	1	1	3	1	0	22	2.59
			Placebo	2	2	1	1	2	3	1	0	4	16	4.25

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; at least Some pain obtained by combining levels 2–4.
[b]Initial pain status reported only as either 2:Some or 3:Lots, despite potential levels of 0–4.

relative to the null hypothesis of "no partial association." For comparative purposes, we present each of these three test statistics in Table 6 adjusted for the eight strata formed by center and initial pain level and without this adjustment. If we treat both the factor levels and response levels as nominal (i.e. ignoring any inherent ordering to the levels), then the randomization model test statistic in expression (A1.6) of Appendix 1 provides a test of the null hypothesis relative to the alternative that there is a general, nonspecific association between treatment regimen and total number of hours with at least

Some (levels 2–4) postpartum pain. The adjusted test statistic is $Q_{MH(1)} = 76.37$, in contrast to the unadjusted test, $Q = 72.78$. Here again, the effect of adjusting for the covariates seems minimal. In both cases, the test statistics with d.f. = 16 are highly significant ($p < 0.01$), suggesting important, nonspecific differences in the level of pain among the treatments.

Suppose now that we incorporate the ordinal level of the response variable into the analysis, but still consider the levels of treatment as nominal. Within each stratum, we can compute an observed mean score (i.e. average number of hours with at least Some postpartum pain) for each treatment regimen, as summarized in the last column of Table 5. For example, in the first row of Table 5, the observed mean score for patients with Some initial pain in Center No. 1 receiving the combination drug (A & B) is obtained as

$$[\,(6\times0) + (4\times1) + (5\times2) + (3\times3) + (1\times4) + (0\times5) + (0\times6) + (0\times7) + (0\times8)\,] / 19 = 1.42.$$

Now we can use the randomization model test statistic in expression (A1.10) of Appendix 1 to test the null hypothesis of "no partial association" against the more specific alternative hypothesis that the mean scores for the three treatment regimens are different, adjusting for the strata. The adjusted test statistic is $Q_{MH(2)} = 58.02$, in contrast to the unadjusted test statistic, $Q = 52.86$. Since these tests compare the three mean scores (in a fashion equivalent to ANOVA), the degrees of freedom have been reduced from 16 to 2, resulting in increased statistical power to reject H_0. Both test statistics are highly significant ($p < 0.01$); moreover, the 2 d.f. mean score test statistic represents 76% of the total variation (58.02 / 76.37) statistic with 16 d.f. Thus, the general, nonspecific association between treatment and number of hours with at least Some postpartum pain largely is due to the more specific differences among mean scores for each treatment.

Table 6 Randomization model test statistics for association between treatment and number of hours of at least some post-partum pain[a] both with and without adjustment for center by initial pain strata

		Effect of adjustment for center by initial pain strata			
		Stratum-adjusted covariance structure		Unadjusted covariance structure	
Alternative hypothesis	d.f.	Test statistic	Signif. level	Test statistic	Signif. level
General association	16	76.37	<0.01	72.78	<0.01
Row mean scores differ	2	58.02	<0.01	52.86	<0.01
Trend in mean scores	1	57.72	<0.01	52.46	<0.01

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; at least Some pain obtained by combining levels 2–4.

Finally, we consider the ordinal dimension in the treatment levels. We might expect a progression in the mean scores from shorter duration of at least Some pain for patients receiving drugs (A & B) to largest for those on placebo. We can investigate this progression in mean scores by computing the test statistic in expression (A1.14) from Appendix 1 using equally spaced row scores of 1, 2, 3. Other methods for assigning row scores, such as those based on ranks, may be preferred by some investigators for this setting to avoid the equal spacing assumption inherent in the integer scoring. The adjusted test statistic is $Q_{MH(3)} = 57.72$, in contrast to the unadjusted test statistic, $Q = 52.46$. With 1 degree of freedom, these test statistics are highly significant ($p < 0.01$). Note that virtually all of the nonspecific difference in mean scores ($Q_{MH(2)} = 58.02$) is encompassed by a linear trend in the mean scores ($Q_{MH(3)} = 57.72$). Thus, these data provide strong evidence that increasing the treatment regimen (from placebo to Drug A alone to Drugs A & B) is associated with a linear decrease in the average number of hours women report at least some postpartum pain.

4. REPEATED MEASURES HYPOTHESES

The use of Mantel-Haenszel methods for the analysis of data obtained from repeated measures designs probably is less well known than for standard factor-response designs. However, several special cases were described in the early papers by Mantel & Haenszel (43), Birch (3, 4), and Gart (24). More recently, Breslow & Day (7) illustrated the use of these methods for matched case-control data, including the incorporation of scores for an ordinal risk factor. Otherwise, Mantel & Byar (41), Darroch (13), and White et al (55) described the use of these methods for various other repeated measures applications.

In this section we illustrate the use of the methods for repeated measures, from the simple (2×2) table situation to the more complex layout of a multilevel ordinal response variable measured repeatedly over time on the same subjects or units such as matched sets. For uniformity throughout this review, we utilize data from the same example introduced in Section 3.2. Consequently, the hypotheses under consideration now involve the extent to which the level of post-partum pain varies across time within treatment subgroups. Although these hypotheses may be of less interest in multicenter clinical trials concerned with treatment efficacy, they are identical to the primary hypotheses in matched case-control studies involving ordinal risk factors such as those illustrated in Breslow & Day (7). In this setting, the matched set (rather than the subject) is the observational unit, so that the specific members of the matched set (such a case, hospital control, neighborhood control, etc) correspond to the row levels of the subtable, and the levels of the risk factor are connected to the columns. In fact, all the matched

analyses in that text, can be formulated within this framework utilizing the Mantel-Haenszel methods.

4.1. (2 × 2) *Tables*

Quite often there is interest in comparing the prevalence of a risk factor in a matched set of subjects or the change in the prevalence of a condition under different circumstances or time periods. In such applications, a 2 × 2 table is constructed either for each matched set or for each subject, depending on the nature of the repeated measure structure. In either case, the row and column dimension correspond to the level of the risk factor or condition for each dimension of the repeated measure. For example, in the study of post-partum pain described previously, the treatments were administered at baseline (hour 0) and after 4 hr of follow-up (hour 4). Although one may anticipate a sizable difference among treatments in the first hours after delivery when pain is likely to be most acute, it is less clear whether any significant differences would be observed between hour 4 and hour 8 (the end of follow-up). Thus, for this post-partum pain example, we may be interested in change in level of pain at eight hours from the level at four hours, utilizing the data for each patient. We can utilize the same statistic presented in Expression 2 to test the null hypothesis that the level of post-partum pain [at least Some (levels 2–4) vs little or None [0–1)] is the same at 4 and 8 hr after delivery.

In order to implement this test with appropriate adjustments for the repeated measures structure within subjects, the data for each individual must be summarized in a (2 × 2) frequency table with entries of 1's and 0's, as shown for illustrative purposes in Table 7 for Patient No. 184 assigned to the placebo treatment in Center No. 1. At hour 4, this patient reported Little or No pain; whereas at hour 8 she reported at least Some pain.

Table 7 Level of post-partum pain[a] at 4 and 8 hours after delivery for Patient No.184 from Center No. 1: Treatment = Placebo; initial pain status = Lots[b]

Number of hours after delivery	Level of post-partum pain		Total
	S	N	
4	0	1	1
8	1	0	1
Total	1	1	2

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; S denotes at least Some pain (levels 2–4) and N denotes little of no pain (levels 0–1).

[b]Initial pain status reported only as either 2:Some or 3:Lots, despite potential levels of 0–4.

In summarizing these data from a set of (2 × 2) tables across all subjects, standard notation for the joint frequencies of the two-level response variable at each of the two conditions or time points commonly is employed. In particular, this general notation for the cross-classification of post-partum pain levels at 4 and 8 hr after delivery is presented in Table 8. This layout is identical to the format used to summarize the frequencies for case-control data from epidemiologic investigations involving matched-pairs, where the rows represent the exposure categories for the cases and the columns represent the exposure categories for the controls.

The joint frequency distribution of the postpartum pain levels at 4 and 8 hr after delivery are summarized by treatment subgroup and initial pain status in Table 9. In this context, the randomization model test statistic in Expression 2 simplifies to $Q_{MH(1)} = (b - c)^2 / (b + c)$, which is identical to the McNemar (45) statistic for repeated measures data from (2 × 2) tables. This test statistic asymptotically follows the chi-square distribution with d.f. = 1. Note, however, that it is completely determined by the off-diagonal cells "*b*" and "*c*", and thus the asymptotics for this statistic are linked directly to the number of discordant pairs of observations, $b + c$, rather than the total sample size. In order to emphasize the importance of these two cells, their values appear in bold print in Table 9. When $b + c$ does not exceed 20, this McNemar statistic may not follow the chi-square distribution adequately; in such situations it is preferable to perform a small-sample test based on the binomial distribution.

In Table 9, we note that the test statistics for women with Some pain at baseline range from 3.5 for the placebo treatment to 4.0 for the conbination drug, suggesting only slight evidence that the level of pain at 8 hr differs from that at 4 hr. On the other hand, for women with initial pain status of Lots, there is a large variation among treatments, ranging from no evidence for a difference in pain level in the Placebo group ($Q_{MH(1)} = 0.47$), to strong evidence for such a difference in the A only group ($Q_{MH(1)} = 7.54$), to very strong evidence in the group receiving both A & B ($Q_{MH(1)} = 13.76$). When

Table 8 General notation for cross-classification of post-partum pain[a] levels at four and eight hours after delivery

Pain at four hours	Pain at eight hours	
	S	N
S	*a*	*b*
N	*c*	*d*

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; S denotes at least Some pain (levels 2–4) and N denotes little of no pain (levels 0–1).

Table 9 Joint frequency distribution, randomization model test statistic and measure of association for level of post-partum pain[a] at four and eight hours after delivery, by treatment and initial pain status

Treatment subgroup	Initial pain	Response profile at hours 4 and 8 SS (a)	SN (b)	NS (c)	NN (d)	Total	$Q_{MH(1)}$ (d.f. = 1)[b]	Signif. level	MH avg. odds ratio
Placebo	Some	17	**16**	**7**	40	80	3.52	0.06	2.29
	Lots	41	**11**	**8**	30	90	0.47	0.49	1.38
	Combined	58	**27**	**15**	70	170	3.43	0.06	1.80
A only	Some	11	**15**	**6**	46	78	3.86	0.05	2.50
	Lots	31	**20**	**6**	41	98	7.54	<0.01	3.33
	Combined	42	**35**	**12**	87	176	11.26	<0.01	2.92
A & B	Some	5	**12**	**4**	64	85	4.00	0.05	3.00
	Lots	14	**19**	**2**	65	100	13.76	<0.01	9.50
	Combined	19	**31**	**6**	129	185	16.89	<0.01	5.17

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; S denotes at least Some pain (levels 2–4) and N denotes little of no pain (levels 0–1).
[b]$Q_{MH(1)}$ is identical to McNemar's (1947) Statstic.

initial pain status is ignored by combining the strata within treatment, the test statistics demonstrate a clear gradient across treatments, with the least evidence for association among women receiving Placebo and the strongest evidence for association among women receiving both A & B. Thus, although the general hypothesis-testing framework for this repeated measures structure does not provide a direct test for treatment differences, formal tests for treatment differences in this situation would involve treatment × time interaction effects. The magnitude of these within-treatment test statistics suggests the same ordering of effectiveness noted previously.

The usual Mantel-Haenszel average odds ratio as defined in Expression 3 simplifies to $\hat{\psi}_{MH} = b / c$ (using the notation of Table 8) for (2 × 2) tables from repeated measures designs. These odds ratios for each combination of treatment subgroup and initial pain status are presented in the last column of Table 9. Note again that these odds ratios for women reporting Some initial pain are similar across treatments, ranging from 2.3 for those receiving Placebo to 3.0 for those receiving both A & B.

Among women who reported an initial pain status of Lots, the large treatment effect suggested by the test statistics also are reflected in the odds ratios. For example, in the Placebo subgroup, the number of women reporting

a decline in pain from hour 4 to hour 8 only slightly exceeded the number reporting an increase ($\hat{\psi}_{MH} = 1.38$). Drug A was associated with a larger effect ($\hat{\psi}_{MH} = 3.33$), but the most impressive results were observed in the A & B subgroup, where 19 of the 21 women who reported different levels of pain experienced a decrease from at least Some at hour 4 to Little or None at hour 8 ($\hat{\psi}_{MH} = 9.50$). When initial pain status is ignored by combining the strata within treatment, the odds ratios demonstrate a clear gradient across treatments, with the least evidence for a difference in level of pain at 4 and 8 hr of follow-up among women receiving Placebo and the strongest evidence for such a difference among women receiving both A & B.

Overall, although the possibility of treatment × time interaction has not been considered, these patterns suggest that for women who report Some pain at baseline there is no discernable difference among treatments for change between hour 4 and hour 8. One implication may be that perhaps only placebo should be offered at hour 4 for these women; whereas for women who report Lots of pain at baseline, the regimen of A & B at hour 4 is the treatment of choice.

Several of the appropriate variance formulae characterized in Table 2 for the Mantel-Haenszel average odds ratio under sampling model 2 can be considered for these average odds ratios. In fact, for these (2 × 2) table

Table 10 Estimated standard errors for ln Mantel-Haenszel (MH) average odds ratio and 95% confidence intervals for MH average odds ratio for level of post-partum pain[a] at four and eight hours after delivery by treatment and initial pain status

Treatment subgroup	Initial pain status	MH avg. odds ratio	Estimated standard error for ln MH avg. odds ratio	95% Confidence interval for MH avg. odds ratio[b]
Placebo	Some	2.29	0.45	(0.94, 5.56)
	Lots	1.38	0.46	(0.55, 3.42)
	Combined	1.80	0.32	(0.96, 3.38)
A only	Some	2.50	0.48	(0.97, 6.44)
	Lots	3.33	0.47	(1.34, 8.30)
	Combined	2.92	0.33	(1.51, 5.62)
A & B	Some	3.00	0.58	(0.97, 9.30)
	Lots	9.50	0.74	(2.21, 40.79)
	Combined	5.17	0.45	(2.16, 12.38)

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; S denotes at least Some pain (levels 2–4) and N denotes little of no pain (levels 0–1).

[b]For matched pairs design, V_{B2}, V_F, and V_R are identical; V_{CS} is nearly identical.

results, there is little difference among these formulae; some are mathematically identical. These were used to compute estimated standard errors for the ln Mantel-Haenszel average odds ratio; 95% confidence intervals for the M-H average odds ratio, computed separately by initial pain status and treatment, are summarized in Table 10.

It is of interest to note that these confidence intervals are not totally consistent with the test statistics presented in Table 9. In particular, the two test statistics for women with Some initial pain in the A Only and A & B treatment subgroups are significant at the 5% level, whereas both lower 95% confidence limits include 1.0. These differences occur largely due to the sparse sample sizes for the "b" and "c" cells.

4.2. *General* (s × 2) *Tables*

In the preceding section we compared the level of postpartum pain (Little or None vs at least Some) at $s = 2$ time points. We now extend that analysis to compare the levels of postpartum pain at each hour after delivery within treatment subgroups. The overall prevalence of at least Some post-partum pain for each of the three treatment subgroups are displayed graphically in Figure 1. Note that the proportion of patients reporting at least Some pain at baseline (hour 0) is the same for all three treatments, reflecting the anticipated effect of randomization to treatment. However, for the 8 hr of follow-up, the proportion of patients reporting at least Some pain is consistently highest for the Placebo group and lowest in the group receiving drugs A & B.

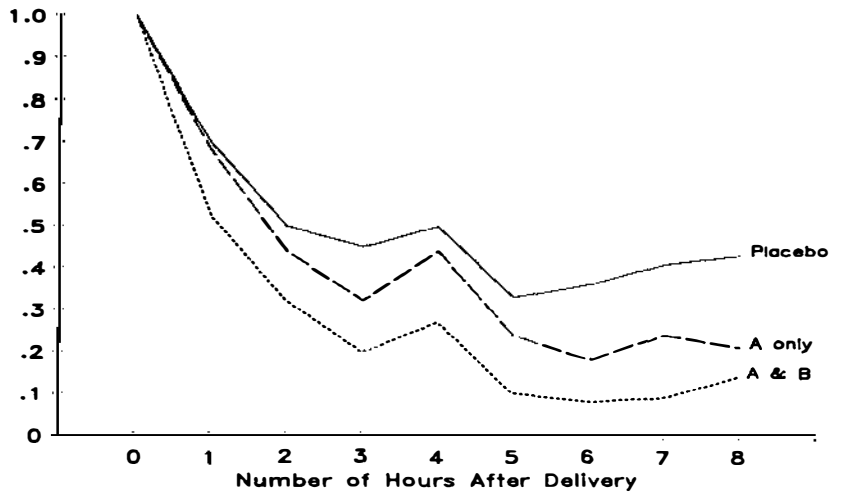

Figure 1 Observed proportion of women with at least Some postpartum pain by number of hours after delivery and treatment.

The statistics presented in Expressions (A1.6), (A1.10), and (A1.14) from Appendix 1 can be used to test the null hypothesis that the level of postpartum pain is not associated with the number of hours after delivery against the three alternative hypotheses defined in Section 2.1–2.3. The number of response levels (columns) for these analyses remains at 2 (at least Some pain vs Little or None); whereas the number of time points (rows) has increased from 2 to 8. In general epidemiological applications, such as matched case-control studies, each of the s rows correspond to a different patient-type within the matched set, e.g. case, hospital control, and neighborhood control, and matched sets correspond to patients.

For illustrative purposes, consider again Patient No. 184 from Center No. 1. The response profile for this patient's level of post-partum pain (Little or None vs at least Some) at each hour of follow-up after delivery is presented in Table 11. Thus, this patient reported at least Some pain for the first 2 hr after delivery, Little or No pain for hours 3, 4, and 5 after delivery, and at least Some pain again for hours 6, 7, and 8 after delivery. The data for each individual must be summarized in similar (8×2) tables, which form the strata necessary to implement the tests of marginal homogeneity of the level of pain across time with the appropriate adjustments for the repeated measures structure within subjects.

In order to illustrate the importance of adjusting for the correlation among measurements within patients, each of the generalized Mantel-Haenszel test statistics were computed both for the unstratified data [one (8×2) table with frequencies collapsed over patients] and across the subtables corresponding to

Table 11 Distribution of level of post-partum pain[a] by number of hours after delivery for Patient No. 184 from Center No. 1: Treatment = Placebo; Initial pain status = Lots[b]

Number of hours after delivery	Level of post-partum pain		Total
	N	S	
1	0	1	1
2	0	1	1
3	1	0	1
4	1	0	1
5	1	0	1
6	0	1	1
7	0	1	1
8	0	1	1

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; S denotes at least Some pain (levels 2–4) and N denotes little of no pain (levels 0–1).

[b]Initial pain status reported only as either 2:Some or 3:Lots, despite potential levels of 0–4.

each of the patients within treatment subgroups. Unlike the previous analyses, which were conducted separately for patients by initial pain status, we present only the combined analyses for purposes of consciseness. The results for the stratified analysis are reported in Table 12 in the panel of columns labeled "Stratum-Adjusted Covariance Structure"; whereas the results from the unstratified analysis are provided in the columns labeled "Unadjusted Covariance Structure." All of these statistics were generated directly within the FREQ procedure from SAS (50), although the stratified analyses do require a bit of special pre-programming to construct the subtables of the type illustrated in Table 11.

In these analyses, we are interested primarily in the test statistic with 1 d.f. directed at the alternative hypothesis that suggests a linear trend in mean scores, although the mean score difference and general association tests with d.f. = 7 are provided for purposes of completeness. Furthermore, the asymptotic approximation to the chi-square distribution is much better for the 1 d.f. statistic. The results presented in Table 12 demonstrate the fact that when the response variable has only two levels, the degrees of freedom and the test statistic for the alternative hypotheses directed at general association and different row mean scores are identical. Each of the test statistics is highly significant ($p<0.01$). This suggests that the level of postpartum pain tends to decrease from at least Some to Little or None with increasing time after

Table 12 Randomization model test statistics for association between time after delivery and response of a least some post-partum pain[a], by treatment, both with and without adjustment for patient strata

			Effect of adjustment for patient strata			
			Stratum-adjusted covariance structure		Unadjusted covariance structure	
Treatment subgroup	Alternative hypothesis	d.f.	Test statistic	Signif. level	Test statistic	Signif. level
Placebo	General association	7	107.25	<0.01	63.02	<0.01
	Row mean scores differ	7	107.25	<0.01	63.02	<0.01
	Trend in mean scores	1	54.34	<0.01	31.93	<0.01
A only	General association	7	212.76	<0.01	139.75	<0.01
	Row mean scores differ	7	212.76	<0.01	139.75	<0.01
	Trend in mean scores	1	125.93	<0.01	82.72	<0.01
A & B	General association	7	234.71	<0.01	179.83	<0.01
	Row mean scores differ	7	234.71	<0.01	179.83	<0.01
	Trend in mean scores	1	166.48	<0.01	127.55	<0.01

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up; at least Some pain obtained by combining levels 2–4.

delivery for each treatment subgroup. Note, however, that there is a distinct progression in the test statistics for this linear trend alternative from $Q_{MH(3)} = 54.34$ for patients receiving Placebo, to 125.93 for patients receiving drug A only and to 166.48 for patients receiving both drugs A & B. Thus, since these groups have similar numbers of patients, the evidence for improvement over time is strongest in the A & B treatment subgroup.

For each of these test statistics, the unadjusted counterpart obtained by ignoring stratification by patient is smaller, reflecting the anticipated loss in power to detect within subgroup differences across repeated measures conditions. The fact that these test statistics also are highly significant ($p<0.01$) should not be interpreted to imply that the more complicated adjustment procedures are unimportant; on the contrary, situations with less substantial time trends could lead to contradictory conclusions if the within-subject correlation is ignored.

4.3. *General* (s × L) *Tables*

In the most general setting for a repeated measures design there are s time points (measurement conditions or matched-set members) and L levels of the response variable. Recall that there are actually five distinct levels of postpartum pain, ranging from 0 (None) to 4 (Terrible), which were recorded for each patient at each hour of follow-up. Thus, it is reasonable to consider extending the same analyses described in the preceding section for only two

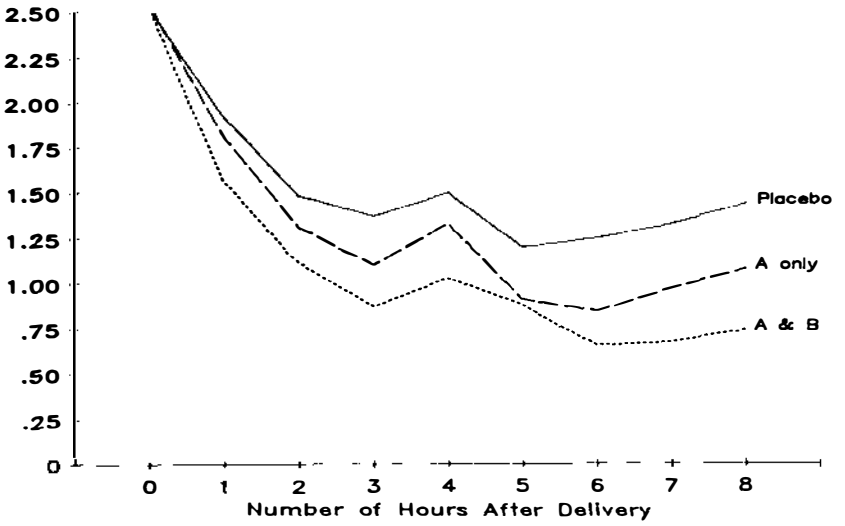

Figure 2 Observed mean score of postpartum pain by number of hours after delivery and treatment.

levels of response to the expanded scale with five levels of response. The overall response profiles for the mean score of the pain level within each treatment regimen are displayed graphically in Figure 2. These mean scores for each of the three treatment groups are the same at baseline (hour 0). However, the largest mean scores are consistently observed among patients in the Placebo group, and the smallest mean scores are consistently observed among patients receiving both drugs A & B.

The null hypothesis of interest is that the level of postpartum pain (0 through 4) is not associated with the number of hours after delivery, with the alternative hypothesis positing that there is a linear decrease in the mean pain score from hour 1 to hour 8. In order to incorporate the appropriate adjustments for the repeated measures across time, the data for each patient must be structured as shown for the representative patient No. 184 from Center No. 1 in Table 13. This patient reported No pain (level 0) at hour 4, Little pain (level 1) at hours 3 and 5, and Some pain (level 2) the remaining hours of follow-up after delivery.

The results for the randomization model test statistics against all three alternative hypotheses within treatment subgroups, both with and without adjustment for the repeated measures covariance structure, are presented in Table 14. Once again, all of these results are highly significant ($p<0.01$). The degrees of freedom for the general association alternative in the A Only and A & B subgroups is 21 rather than 28 (as in the Placebo subgroup), since none of the patients in these two treatment regimens reported a pain level of 4 (Terrible). Within each treatment, the data strongly suggest a linear decrease in mean pain score with increasing time after delivery. As before, however,

Table 13 Distribution of level of post-partum pain[a] by number of hours after delivery for Patient No. 184 from Center No. 1: Treatment = placebo; initial pain status = Lots[b]

Number of hours after delivery	Level of post-partum pain					
	0	1	2	3	4	Total
1	0	0	1	0	0	1
2	0	0	1	0	0	1
3	0	1	0	0	0	1
4	1	0	0	0	0	1
5	0	1	0	0	0	1
6	0	0	1	0	0	1
7	0	0	1	0	0	1
8	0	0	1	0	0	1

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up.

[b]Initial pain status reported only as either 2:Some or 3:Lots, despite potential levels of 0–4.

there is a distinct progression in the value of these test statistics across treatments, ranging from $Q_{MH(3)} = 60.45$ for the Placebo regimen, to $Q_{MH(3)} = 153.22$ for the A Only regimen, to $Q_{MH(3)} = 230.12$ for the A & B regimen. Each of these values is larger than their counterparts in Table 12, where level of pain was dichotomized. Thus, this most general setting for repeated measures data provides the strongest evidence that receiving both drugs A & B is the most effective treatment for reducing the level of pain in the first 8 hr after delivery.

CONCLUDING REMARKS

We have made no attempt to be exhaustive in our coverage of all the types of research problems for which the Mantel-Haenszel methodology has been applied successfully. For example, the Mantel-Haenszel test statistic $Q_{MH(1)}$ also has been used widely in the analysis of survival data. In this setting, the duration of time from baseline, e.g. the initiation of a cancer treatment, to a target event, such as death, is recorded for each subject in the study. Quite often, the study involves subjects who have been randomized into several treatment groups. These data can be summarized in an ($s \times 2$) contingency table, where the "s" treatment groups form the rows and the two response outcomes (dead vs alive) form the columns at each distinct time point for which an event occurred in any of the treatment groups. Thus, there will be as many subtables as these are total number of event, e.g. deaths, in the study.

Table 14 Randomization model test statistics for association between time after delivery and level of lost-partum pain[a], by treatment, both with and without adjustment for patient strata

			Effect of adjustment for patient strata			
			Stratum-adjusted covariance structure		Unadjusted covariance structure	
Treatment subgroup	Alternative hypothesis	d.f.	Test statistic	Signif. level	Test statistic	Signif. level
Placebo	General association	28	205.22	<0.01	104.54	<0.01
	Row mean scores differ	7	156.65	<0.01	72.68	<0.01
	Trend in mean scores	1	60.45	<0.01	28.05	<0.01
A only	General association	21	298.86	<0.01	168.17	<0.01
	Row mean scores differ	7	270.00	<0.01	140.59	<0.01
	Trend in mean scores	1	153.22	<0.01	79.78	<0.01
A & B	General association	21	360.17	<0.01	236.45	<0.01
	Row mean scores differ	7	328.30	<0.01	196.20	<0.01
	Trend in mean scores	1	230.12	<0.01	137.52	<0.01

[a]Pain level reported as 0:None, 1:Little, 2:Some, 3:Lots, 4:Terrible at each hour of follow-up.

Mantel (39a) proposed using $Q_{MH(1)}$ to test the hypothesis that, at each point in time, the risk (or hazard) of death is the same for each treatment. In this context, $Q_{MH(1)}$ is known as the logrank test, and is used to compare the survival curves of the different groups. Many of the other test statistics that have been proposed for this purpose are actually generalizations of the logrank test. However, when the proportional hazards model holds (implying a constant relative risk at each time point), then $Q_{MH(1)}$ is the best test statistic in the large class of logrank tests. Thus, Mantel and Haenszel's logrank test has become the standard in cancer clinical trials and other survival analysis applications. More detail on the logrank test and its extensions are provided in Crowley et al (12a).

It is a measure of the genius of these procedures that they continue to be useful in a wide-ranging set of problems, including survival analysis, epidemiological studies with matched designs, and observational and historical studies for which a minimum of assumptions about the target population are tenable. We have attempted to illustrate the enormous flexibility of this very general methodology within the context of the same data set. Undoubtedly further extensions and applications of the Mantel-Haenszel methodology will continue to appear in the scientific literature.

APPENDIX 1

Matrix Formulations for the Mantel-Haenszel Test Statistics

For each table, as in Table 1, let $\mathbf{P}_{hi\cdot} = (N_{hi\cdot}/N_h)$ denote the marginal row proportion in the hth stratum and let $\mathbf{P}_{h\cdot j} = (N_{h\cdot j}/N_h)$ denote the marginal column proportion in the hth stratum. These proportions can be summarized in vector notation as $\mathbf{P}_{h*\cdot}$ and $\mathbf{P}_{h\cdot *}$, respectively. Under the hypergeometric model defined in Eq. 1, the expected value of $\mathbf{n}_h$ is

$$\mathbf{m}_h = N_h[\mathbf{P}_{h\cdot *} \otimes \mathbf{P}_{h*\cdot}], \qquad \text{A1.1.}$$

where $\otimes$ denotes left-hand Kronecker product multiplication, the matrix on the left of the $\otimes$ symbol being multiplied by each element in the matrix on the right. Similarly, the covariance matrix of $\boldsymbol{n}_h$ under H_0 is

$$V_h = \frac{N_h^2}{(N_h - 1)} (\boldsymbol{D}_{\mathbf{P}_{h\cdot *}} - \mathbf{P}_{h\cdot *}\mathbf{P}_{h\cdot *}') \otimes (\boldsymbol{D}_{\mathbf{P}_{h*\cdot}} - \mathbf{P}_{h*\cdot}\mathbf{P}_{h*\cdot}'), \qquad \text{A1.2.}$$

where $\boldsymbol{D}_{\boldsymbol{P}_h}$ is a diagonal matrix with elements of the vector $\boldsymbol{P}_h$ on the main diagonal.

Without loss of generality, let $\mathbf{A}_1 = [(\mathbf{I}_{(r-1)}, \mathbf{0}_{(r-1)}) \otimes (\mathbf{I}_{(s-1)}, \mathbf{0}_{(s-1)})]$ be a linear operator matrix, which eliminates the last row and column frequencies from each stratum arrayed as in Table 1, and let

$$\mathbf{G}_h = \mathbf{A}_1(\mathbf{n}_h - \mathbf{m}_h) \qquad \text{A1.3.}$$

represent the differences between the observed and expected frequencies under H_0 for the $u = (s - 1)(r - 1)$ pivot cells. Then it follows that $E\{\mathbf{G}_h|H_0\} = \mathbf{0}_u$ and $\mathbf{Var}\{\mathbf{G}_h|H_0\} = \mathbf{A}_1\mathbf{V}_h\mathbf{A}_1'$.

Now if we let

$$\mathbf{G} = \sum_{h=1}^{t} \mathbf{G}_h \qquad \text{A1.4.}$$

be the sum of these differences across the t strata and

$$\mathbf{V}_\mathbf{G} = \sum_{h=1}^{t} \mathbf{Var}\{\mathbf{G}_h|H_0\} \qquad \text{A1.5.}$$

be the corresponding sum of the covariance matrices, then the generalized Mantel-Haenszel test for H_0 against the nonspecific alternative that the marginal distributions of the response variable are not homogeneous across the s subpopulations is

$$Q_{\mathrm{MH}(1)} = \mathbf{G}'\mathbf{V}_\mathbf{G}^{-1}\mathbf{G}, \qquad \text{A1.6.}$$

which follows a chi-square distribution with d.f. $= u$ under H_0, provided $N = \Sigma N_h$ is large and expected values satisfy the Mantel-Fleiss criteria (42).

If the response variable is reported on an ordinal scale, let $\mathbf{A}_{2h}$ be a $s \times sr$ block-diagonal matrix, with respective blocks being $\mathbf{a}_h' = (a_{h1}, \ldots, a_{hr})$, where a_{hj} is an appropriate score reflecting the ordinal nature of the jth level of response for the hth stratum. Then it follows that

$$\mathbf{M}_h = \mathbf{A}_{2h}(\mathbf{n}_h - \mathbf{m}_h) \qquad \text{A1.7.}$$

represents the differences between the observed and expected weighted mean scores under H_0 for the s subpopulations for the hth stratum. Then it follows that $E\{\mathbf{M}_h|H_0\} = \mathbf{0}_s$ and the covariance matrix for these correlated mean scores can be written as $\mathbf{Var}\{\mathbf{M}_h|H_0\} = \mathbf{A}_{2h}\mathbf{V}_h\mathbf{A}_{2h}'$. Now if we let

$$\mathbf{M} = \sum_{h=1}^{t} \mathbf{M}_h \qquad \text{A1.8.}$$

be the sum of these weighted mean score differences across the t strata,

$$\boldsymbol{V}_{\mathbf{M}} = \sum_{h=1}^{t} \boldsymbol{Var}\{\mathbf{M}_h|H_0\} \qquad \text{A1.9.}$$

be the corresponding sum of the subject-specific covariance matrices, and $\boldsymbol{R}$ be any contrast matrix of rank $(s - 1)$ that spans the space of the s mean scores, then it follows that

$$Q_{\text{MH}(2)} = (\boldsymbol{R}\mathbf{M})'[\boldsymbol{R}\boldsymbol{V}_{\mathbf{M}}\boldsymbol{R}']^{-1}\boldsymbol{R}\mathbf{M} \qquad \text{A1.10.}$$

provides a test for the equality among the mean responses for the s subpopulations relative to the response variable score vectors $\{\boldsymbol{a}_h\}$. Under H_0, it can be shown that $Q_{\text{MH}(2)}$ follows the chi-square distribution with d.f. $= s - 1$ for large N.

If we let $\mathbf{c}_h' = (c_{h1}, \ldots, c_{hs})$, where c_{hi} is an appropriate score for the ith factor level from the hth stratum, then it is possible to construct a statistic directed at the extent to which the mean response varies linearly across levels of the factor. For this purpose, let $\mathbf{A}_{3h} = \boldsymbol{a}_h' \otimes \mathbf{c}'$ be the $(1 \times sr)$ vector with elements formed by the product of the corresponding row and column ordinal scores. Then it follows that the scalar

$$C_h = \mathbf{A}_{3h}(\mathbf{n}_h - \mathbf{m}_h) \qquad \text{A1.11.}$$

represents the difference between the observed and expected weighted sum of the product of the row and column scores under H_0 for the hth stratum, and as such, is directed at linear correlation alternative hypotheses. The variance of this linear statistic can be expressed directly as $\boldsymbol{Var}\{C_h|H_0\} = \mathbf{A}_{3h}\boldsymbol{V}_h\mathbf{A}'_{3h}$.

Now if we let

$$C = \sum_{h=1}^{t} C_h \qquad \text{A1.12.}$$

be the sum of these linear combinations of scores across the t strata,

$$\boldsymbol{V}_C = \sum_{h=1}^{t} \boldsymbol{Var}\{C_h|H_0\} \qquad \text{A1.13.}$$

be the corresponding sum of the stratum-specific variances, then it follows that

$$Q_{MH(3)} = \frac{C^2}{\boldsymbol{V}_C} \qquad \text{A1.14.}$$

provides a test for the linear trend in the mean response across the subpopulations relative to the score vectors $\{\mathbf{c}_h\}$ and the $\{\mathbf{a}_h\}$. Under H_0, it can be shown that $Q_{MH(3)}$ follows the chi-square distribution with d.f. = 1 for large N.

APPENDIX 2

Variance Formulae for the Mantel-Haenszel Average Odds Ratio

For notational unification, let

$$W_h = (var\{\log_e(\hat{\psi}_h)\})^{-1}$$
$$= (n_{h11}^{-1} + n_{h21}^{-1} + n_{h12}^{-1} + n_{h22}^{-1})^{-1};$$

$$u_{1h} = \frac{n_{h11}\, n_{h21}\, n_{h\cdot 2}}{n_{h11}\, n_{h21}\, n_{h\cdot 2} + n_{h12}\, n_{h22}\, n_{h\cdot 1}};$$

$$u_{2h} = \frac{n_{h12}\, n_{h22}\, n_{h\cdot 1}}{n_{h11}\, n_{h21}\, n_{h\cdot 2} + n_{h12}\, n_{h22}\, n_{h\cdot 1}}.$$

Also, let R_h and S_h denote the contribution from each stratum to the numerator and denominator, respectively, of the Mantel-Haenszel average odds ratio, so that

$$R_h = n_{h11}n_{h22}/N_h;$$
$$S_h = n_{h12}n_{h21}/N_h;$$
$$\hat{\psi}_{\mathrm{MH}} = \sum_{h=1}^{t} R_h / \sum_{h=1}^{t} S_h.$$

Then the various formulae for the variance of the Mantel-Haenszel odds ratio, appropriate only under standard large-sample theory, are defined below.

$$v_{\mathrm{H}} = (\hat{\psi}^2_{\mathrm{MH}}) \left[\frac{\sum_h S_h^2/W_h}{\left(\sum_h S_h\right)^2} \right]; \qquad \text{A2.1.}$$

$$v_{HS} = (\hat{\bar{\psi}}^2_{MH})\left[\frac{\sum_h (S_h^2/W_h)\ \sum_h (R_h^2/W_h)}{\left(\sum_h S_h\right)^2\left(\sum_h R_h\right)^2}\right]^{1/2}; \qquad \text{A2.2.}$$

$$v_{G1} = \frac{\sum_{h=1}^{t} S_h^2\left[\frac{(n_{h21}\hat{\psi}_h + n_{h11}\hat{\bar{\psi}}_{MH})^2}{n_{h\cdot 1}n_{h11}n_{h21}} + \frac{(n_{h12}\hat{\psi}_h + n_{h22}\hat{\bar{\psi}}_{MH})^2}{n_{h\cdot 2}n_{h12}n_{h22}}\right]}{\left(\sum_{h=1}^{t} S_h\right)^2}; \qquad \text{A2.3.}$$

$$v_{G2} = (\hat{\bar{\psi}}^2_{MH})\ (v_H)\ [(u_{1h}/n_{h\cdot 1})(n_{h21}\hat{\psi}_h + n_{h11}\hat{\bar{\psi}}_{MH}) + (u_{2h}/n_{h\cdot 2})(n_{h12}\hat{\psi}_h + n_{h22}\hat{\bar{\psi}}_{MH})]^2; \qquad \text{A2.4.}$$

$$v_{B2} = \frac{\sum_{h=1}^{t} (R_h - \hat{\bar{\psi}}_{MH}S_h)^2}{\left(\sum_{h=1}^{t} S_h\right)^2}; \qquad \text{A2.5.}$$

$$v_C = \frac{N v_H + t^2 v_{B2}}{N + t^2}. \qquad \text{A2.6.}$$

The variance formula proposed by Flanders (denoted here by v_F) is essentially a modification of v_H in order to be consistent in both limiting models. We can rewrite v_H (which, for a common ψ, is identical to v_{G1} and v_{G2}) as

$$v_H = \frac{(\hat{\bar{\psi}}^2_{MH})\sum_{h=1}^{t} \frac{S_h}{N_h}\left[\frac{(n_{h11} + n_{h22})}{\hat{\psi}_h} + (n_{h21} + n_{h12})\right]}{\left(\sum_{h=1}^{t} S_h\right)^2}. \qquad \text{A2.7.}$$

Now, if we replace $\hat{\psi}_h$ with $\hat{\psi}_{MH}$, add 1 to that numerator term, and subtract 1 from the last term, we obtain the originally proposed v_F as

$$v_F = \frac{(\hat{\psi}^2_{MH}) \sum_{h=1}^{t} \frac{S_h}{N_h} \left[\frac{(n_{h11} + n_{h22} + 1)}{\hat{\psi}_{MH}} + (n_{h21} + n_{h12} - 1) \right]}{\left(\sum_{h=1}^{t} S_h \right)^2}. \quad \text{A2.8.}$$

An alternate form of this variance,

$$v_F = \frac{(\hat{\psi}^2_{MH}) \sum_{h=1}^{t} \left[\frac{S_h}{N_h} \left[\frac{n_{h22}}{\hat{\psi}_{MH}} + n_{h12} \right] + \frac{S_h}{N_h} \left[\frac{(n_{h11} + 1)}{\hat{\psi}_{MH}} + (n_{h21} - 1) \right] \right]}{\left(\sum_{h=1}^{t} S_h \right)^2}$$

A2.9.

is worth noting for purposes of comparison with the subsequent result proposed recently in Robins et al (49).

A different estimator that is also consistent in both limiting models was proposed by Robins et al (49), and is denoted here by v_R. After some algebraic manipulation, it is possible to write v_R as

$$v_R = \frac{(\hat{\psi}^2_{MH}) \sum_{h=1}^{t} \left[\frac{S_h}{N_h} \left[\frac{n_{h22}}{\hat{\psi}_{MH}} + n_{h12} \right] + \frac{R_h}{N_h} \left[\frac{n_{h21}}{\hat{\psi}_{MH}} + \frac{n_{h11}}{\hat{\psi}^2_{MH}} \right] \right]}{\left(\sum_{h=1}^{t} S_h \right)^2}$$

A2.10.

One can readily observe that v_R and v_F are very similar, differing only in the right-hand term of the numerator. In fact, it can be shown that for matched sets with one case per stratum, the difference between v_F and v_R can be expressed as

$$v_F - v_R = \frac{(\hat{\psi}^2_{MH}) \sum_{h=1}^{t} \frac{1}{N_h} [R_h - \hat{\psi}_{MH} S_h]}{\left(\sum_{h=1}^{t} S_h \right)^2}. \qquad \text{A2.11.}$$

When the number of controls matched to each case is constant, then N_h is constant for all h, and v_F is identical to v_R (and both are identical to v_{B2}). Robins et al (49) also proposed a symmetrized version of v_R (denoted here by v_{RS}), where v_{RS} is the arithmetic average of v_R computed first on the original tables, and then recomputed after interchanging the rows of each table.

Literature Cited

1. Agresti, A. 1980. Generalized odds ratios for ordinal data. *Biometrics* 36:59–67
2. Bailar, J. C., Anthony, G. P. 1977. Most cited papers of the *Journal of the National Cancer Institute*, 1962–1975. *J. Natl. Cancer Inst.* 59:709–14
3. Birch, M. M. 1964. The detection of partial association, I: The 2 × 2 case. *J. R. Stat. Soc. Ser. B* 26:313–24
4. Birch, M. W. 1965. The detection of partial association, II: The general case. *J. R. Stat. Soc. Ser. B* 27:111–24
5. Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. 1975. *Discrete Multivariate Analysis*. Cambridge: MIT Press
6. Breslow, N. E. 1981. Odds ratio estimators when the data are sparse. *Biometrika* 68:73–84
7. Breslow, N. E., Day, N. E. 1980. *Statistical Methods in Cancer Research*. Volume 1. *The Analysis of Case-control Studies*. IARC Sci. Publ. No. 32. Lyon: Int. Agency for Res. on Cancer
8. Breslow, N. E., Liang, K. Y. 1982. The variance of the Mantel-Haenszel estimator. *Biometrics* 38:943–52
9. Clayton, D. G. 1974. Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika* 61:525–31
10. Cochran, W. G. 1950. The comparison of percentages in matched samples. *Biometrika* 37:256–66
11. Cochran, W. G. 1954. Some methods for strengthening the common X^2 tests. *Biometrics* 10:417–57
12. Connett, J., Ejigou, A., McHugh, R., Breslow, N. 1982. The precision of the Mantel-Haenszel estimator in case-control studies with multiple matching. *Am. J. Epidemiol.* 116:875–77

12a. Crowley, J., Breslow, N. 1984. Statistical analysis of survival data. *Ann. Rev. Public Health* 5:385–411
13. Darroch, J. N. 1981. The Mantel-Haenszel test and tests of marginal symmetry: Fixed effects and mixed models for a categorical response. *Int. Stat. Rev.* 49:285–307
14. Davis, L. J. 1985. Generalization of the Mantel-Haenszel estimator to nonconstant odds ratios. *Biometrics* 41:487–95
15. Dayal, H. H. 1978. On the desirability of the Mantel-Haenszel summary measure in case-control studies of multifactor etiology of disease. *Am. J. Epidemiol.* 108:506–11
16. Ejigou, A., McHugh, R. 1977. Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics* 33:552–56
17. Flanders, W. D. 1985. A new variance estimator for the Mantel-Haenszel odds ratio. *Biometrics* 41:637–42
18. Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*. New York: Wiley. 2nd ed.
19. Fleiss, J. L. 1984. The Mantel-Haenszel estimator in case-control studies with varying number of controls matched to each case. *Am. J. Epid.* 120:943–52
20. Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32:675–701
21. Gart, J. J. 1962. On the combination of relative risks. *Biometrics* 18:601–10
22. Gart, J. J. 1966. Alternative analyses of contingency tables. *J. R. Stat. Soc. Ser. B* 28:164–79
23. Gart, J. J. 1970. Point and interval estimation of the common odds ratio in the

combination of 2 × 2 tables with fixed marginals. *Biometrics* 57:471–75

24. Gart, J. J. 1971. The comparison of proportions: A review of significance tests, confidence intervals and adjustment for stratification. *Rev. Int. Stat. Inst.* 39: 148–61
25. Gart, J. J., Zweifel, J. R. 1967. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 54:181–87
26. Gilbaud, O. 1983. On the large-sample distribution of the Mantel-Haenszel odds-ratio estimator. *Biometrics* 39: 523–25
27. Greenland, S. 1982. Interpretation and estimation of summary ratios under heterogeneity. *Stat. Med.* 1:217–27
28. Hauck, W. W. 1979. The large-sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* 35:817–19
29. Hauck, W. W., Anderson, S., Leahy, F. J. 1982. Finite-sample properties of some old and some new estimators of a common odds ratio from multiple 2 × 2 tables. *J. Am. Stat. Assoc.* 77:145–52
30. Kleinbaum, D. G., Kupper, L. L., Morgenstern, H. 1982. *Epidemiologic Research.* Belmont: Lifetime Learning Publ.
31. Koch, G. G., Amara, I. A., Davis, G. W., Gillings, D. B. 1982. A review of some statistical methods for covariance analysis of categorical data. *Biometrics* 38:563–95
32. Koch, G. G., Gillings, D. B., Stokes, M. E. 1980. Biostatistical implications of design, sampling, and measurement to health science data analysis. *Ann. Rev. Public Health* 1:163–225
33. Koch, G. G., Imrey, P. B., Singer, J. M., Atkinson, S. S., Stokes, M. E. 1985. *Analysis of Categorical Data.* Montreal: Presses de L'Univ. Montreal
34. Koch, G. G., Edwards, S. 1987. Clinical efficacy trials with categorical data. In *Statistical Methods in the Pharmaceutical Industry,* ed. K. Peace, Chapt. 9. New York: Marcell Dekker. In press
35. Landis, J. R., Heyman, E. R., Koch, G. G. 1978. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Int. Stat. Rev.* 46:237–54
36. Landis, J. R., Cooper, M. M., Kennedy, T., Koch, G. G. 1979. A computer program for testing average partial association in three-way contingency tables (PARCAT). *Comp. Prog. Biomed.* 9:223–46

36a. Lehmann, E. L., Dabrerd, H. J. M. 1975. *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day

37. Liang, K. Y. 1987. Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics* 43:289–99
38. Madansky, A. 1963. Tests of homogeneity for correlated samples. *J. Am. Stat. Assoc.* 58:97–119
39. Mantel, N. 1963. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.* 58:690–700

39a. Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo. Rep.* 50:163–70

40. Mantel, N. 1977. Tests and limits for the common odds ratio of several 2 × 2 contingency tables: Methods in analogy with the Mantel-Haenszel procedure. *J. Stat. Plan. Inf.* 1:179–89
41. Mantel, N., Byar, D. P. 1978. Marginal homogeneity, symmetry and independence. *Commun. Statist.-Theor. Meth. A* 7:953–76
42. Mantel, N., Fleiss, J. L. 1980. Minimum expected cell size requirements for the Mantel-Haenszel one degree of freedom chi-square test and a related rapid procedure. *Am. J. Epidemiol.* 112:129–34
43. Mantel, N., Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* 22:719–48
44. McKinlay, S. M. 1978. The effect of nonzero second-order interaction on combined estimators of the odds ratio. *Biometrika* 65:191–202
45. McNemar, Q. 1974. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12:153–57
46. Mickey, R. M., Elashoff, R. M. 1985. A generalization of the Mantel-Haenszel estimator of partial association for 2 × J × K tables. *Biometrics* 41:623–35
47. Miettinen, O. S. 1976. Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* 103:226–35
48. Phillips, A., Holland, P. W. 1987. Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics* 43:425–31
49. Robins, J., Breslow, N., Greenland, S. 1986. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata limiting models. *Biometrics* 42:311–24
50. SAS Inst., Inc. 1985. *SAS User's Guide: Statistics, 1985 Edition.* Cary, NC: SAS Inst., Inc.

51. Schlesselman, J. J. 1982. *Case-Control Studies*. New York: Oxford Univ. Press
52. Somes, G. W. 1986. The generalized Mantel-Haenszel statistic. *Am. Stat.* 40:106–8
53. Tarone, R. E., Gart, J. J., Hauck, W. W. 1983. On the asymptotic inefficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika* 70:519–22
53a. Thomas, D. G. 1975. Exact and asymptotic methods for the combination of 2 × 2 tables. *Comp. Biomed. Res.* 8:423–46
54. Ury, H. K. 1982. Hauck's approximate large-sample variance of the Mantel-Haenszel estimator. *Biometrics* 38:-1094–95
54a. van Elteren, P. H. 1960. On the combination of independent two-sample tests of Wilcoxon. *Bull. Inst. Intern. Statist.* 37:351–61
55. White, A. A., Landis, J. R., Cooper, M. M. 1982. A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Int. Stat. Rev.* 50:27–34
56. Woolf, B. 1955. On estimating the relation between blood group and disease. *Ann. Hum. Genet.* 19:251–53