On pooled and stratified analysis of epidemiologic rates

Markku Nurminen, Dr PH, PhD

ABSTRACT ......................................................... 164
PREVIEW OF OPTIONS FOR COMBINING DATA .......... 164
POOLABILITY OF RESULTS ................................. 165
INCONSISTENCY OF POOLED ESTIMATORS ............. 167
PREVAILING ANALYTIC PRACTICES FOR
STRATIFIED DATA ............................................. 169
Principles, procedures, and products of data analysis:
incoherent approaches ....................................... 169
Test statistics .................................................. 170
Point estimators .............................................. 173
Interval estimators .......................................... 175
MODEL-BASED APPROACH ................................. 177
The basic approach ........................................ 177
Chi-square function ....................................... 179
ILLUSTRATION ............................................... 181
REFERENCES ............................................... 184
APPENDIX .................................................. 187

Department of Epidemiology and Biostatistics, Institute of Occupational Health,
Topeliuksenkatu 41 a, SF-00250 Helsinki, Finland.
Based on a paper presented to the Working Group on Pooling Strategies in
Occupational Epidemiology of the International Commission on Occupational Health
in connection with the Fourth International Symposium on Epidemiology in

© 1988 Public Health Reviews
Printed in Israel
ABSTRACT
The main object of this paper is to discuss the foundation of methods applicable for the combination of data from several 2×2 contingency tables. The conditions for either a pooled or stratified analysis are examined with a stress on confounding bias. The procedures currently in use are criticized as producing often incoherent and poor results. Specifically, the principles underlying significance testing, and point and interval estimation are reviewed with reference to the value of the epidemiologic relation parameter of interest, either rate difference, rate ratio, or odds ratio. The theoretical basis of a recently suggested approach (Miettinen & Nurminen, 1985) to remedy the inaccuracies of the available data-analytic methods is then briefly outlined. The derived chi square function is based on a model relating the two compared rates. The procedure is recognized as being a generalization of the famous Cochran–Mantel–Haenszel statistic to which it reduces on the null hypothesis for all three alternative parameters. An empirical example illustrates the impact of the proposed improvements on statistical inference.

PREVIEW OF OPTIONS FOR COMBINING DATA
A single nonexperimental study only infrequently provides sufficient data to resolve the presented problem in totality. Thus, there are two basic options available for an epidemiologist for amalgamating his or her study results with the relevant experience obtained from related studies. The researcher can either initially design the study so that it is conducted in a like manner with other investigations sharing common characteristics, or, alternatively, the information yield of independent studies can be combined in the stage of data analysis and inference. The premises for integration of epidemiologic evidence from different sources can best be met in a collaborative project with a unified study protocol, etc. Unfortunately, such joint research undertakings are often infeasible in the context of occupational health research for reasons of high cost and practical hindrances in the coordination. Yet in the case of a rare illness the aggregation of population-experiences may be the only possibility to enhance the reliability of the study to a level where its precision of quantifying
illness rates is reasonably good (see, e.g., Hernberg et al., 1983). When this aggregation is carried out directly, that is, without reference to the original population segment of the study base, the strategy is referred to as pooling. In other words, the occurrence of the illness in question is examined without recognition of the sample from which the disease cases (and noncases) are derived. An alternative procedure is to collect the information from the samples separately and then condition the synthesis of the exposure–illness relationship on the sample-identifying variate.

The conditioning increases precision as just noted, but it may also be used to decrease bias arising from disturbing factors (confounders). In stratification analysis the data are first divided into separate classes according to the possible confounding factors and then the information is accumulated over the strata. (When the stratification is done in the design stage of the study so that the compared samples are selected similarly with respect to the a priori confounding factors the procedure is called matching.)

Alternative means of conditional analysis include restriction and modeling. Restriction is achieved by confinement of the analysis to a homogeneous subdomain of the study population. The modeling approach implies that the (possibly compound) confounding variate is represented in the regression model so that its confounding effect can be quantified. The following presentation is concerned only with the principles and procedures of stratified analysis and pooled (or unstratified) analysis.

POOLABILITY OF RESULTS

Contrary to dose–response experiments, in epidemiology there is no well-founded theory of how to combine the data. Cochran (1954) realized intuitively that the procedure of pooling is legitimate only if the probability of an illness occurrence can be assumed equal. The development of the principles of pooling all the data into a single 2×2 table translates into the consideration of the conditions on which the merger of strata does not introduce confounding.

There are disagreements in the literature about the conceptual definition of confounding and the operational criteria to be used to decide whether or not a measure of exposure–effect relation is biased.
(see, eg., Miettinen & Cook, 1981; Kleinbaum et al., 1982; Boivin & Wachholder, 1985). The essential point is that the expected frequency of occurrence should ideally be the same if the strata are to be pooled (Miettinen, 1976a). Furthermore, this equality should obtain on the null hypothesis, that is, for example, the rates of illness among the nonexposed are to be identical between the strata; this demand is needed because a difference among the exposed may be caused by effect-modification instead of confounding (Miettinen, 1974). Therefore a sufficient condition for a pooled analysis is that the stratification factor in the study base either (a) is not predictive of the illness outcome (a priori) or (b) is unrelated to the determinant exposure under study (cf. Miettinen & Cook, 1981).

In the comprehensive assessment of confounding, complicating problems are often encountered in practice:

- confounders may be determinants of disease diagnosis or they can be involved in the selection of cases and referents (Miettinen & Cook, 1981);
- joint confounding by several extraneous factors (Kleinbaum et al., 1981);
- a sufficient reduction of the list of all potential confounders without too large a residual (uncontrolled) confounding (Cochran, 1965);
- construction of precise operational variates for the confounders (Miettinen, 1983; Nurminen, 1983).

For a detailed discussion of these important and subtle issues of bias, see also Anderson et al. (1980).

In addition to the level-of-principle provisions for confounding, it is commonplace to compute data-based criteria, such as a change-in-estimate of the relation parameter. A usable criterion is provided by the comparison of the ratio of observed to expected counts of exposed cases of illness computed over stratified tables to that obtained from a pooled 2×2 table with the expected numbers derived from the rate among the nonexposed (Miettinen & Cook, 1981; Grayson, 1987). However, the many different approaches that have been suggested (e.g., Rothman, 1975; Miettinen, 1976b; Breslow & Day, 1980; Boivin & Wachholder, 1985) may lead to differing or even false conclusions about the presence of confounding. The propriety of these empirical criteria with regard to the relation parameter at issue
in a particular study design is nowadays better understood but not completely settled (see Kleinbaum et al., 1982; Greenland & Robins, 1985, 1986). Thus it is feasible to use purely statistical criteria together with a priori considerations in an attempt to estimate the extent of the biasing effect of confounding.

In a wider context, namely in general contingency table analysis, rigorous "collapsibility theorems" (Bishop et al., 1975) have been posited for pooling categories without changing the structure of the table. However, the description needs to be formulated in terms of a multivariate (e.g., log-linear) model, and the matter will not be pursued further here.

Given that the conditions for pooling reviewed above hold, we shall proceed to consider a certain statistical property of the pooled estimators.

INCONSISTENCY OF POOLED ESTIMATORS

As pooled analysis is quite generally more efficient than stratified analysis when such pooling leads to a valid estimate of the relative risk (see, e.g., Choi & Howe, 1984), it is important to examine theoretical conditions under which an unbiased estimate of an effect parameter, such as rate ratio (RR), can be obtained.

Consider the combination of a number \( J \) of independent \( 2 \times 2 \) contingency tables. A pooled estimator of RR is computed by adding the respective entries in the tables to obtain

\[
\hat{RR} = \frac{\sum c_{ij}}{\sum S_{ij}} / \left( \frac{\sum C_{ij}}{\sum S_{ij}} \right),
\]

where \( c_{ij} \) and \( S_{ij} \) are the number of cases and subjects in the \( j \)th stratum with the index \( i = 1 \) for the exposed and \( i = 0 \) for the nonexposed individuals, respectively. This estimator, though natural, is problematic in that it gives biased (off-target) values even in large samples because of its inconsistency. To explain this we recast Expression 1 as (cf. Gart, 1962)

\[
\hat{RR} = \frac{\sum S_{ij} \left( c_{ij}/S_{ij} \right) \sum S_{ij}}{\sum S_{ij} \sum S_{ij} \left( c_{ij}/S_{ij} \right)}
\]
\[
\frac{\sum (S_{ij} / \sum S_{ij}) (c_{ij} / S_{ij})}{\sum (S_{0j} / \sum S_{0j}) (c_{0j} / S_{0j})}
= \frac{\sum (S_{ij} / \sum S_{ij}) r_{ij}}{\sum (S_{0j} / \sum S_{0j}) r_{0j}},
\]

(1')

where \( S_i = \sum S_{ij} \) and the (binomial) rate \( r_0 = c_{ij} / S_{ij} \). Now suppose that the sample sizes increase unlimitedly in fixed proportions in such a way that the \( S_i / S_i \)'s approach finite positive constants, \( k_{ij} \) \((i = 1, 0; j = 1, \ldots, J)\), fulfilling the condition \( \sum k_{ij} = 1 \).

Observing that the empirical (sample) rates \( r_{ij} \) are maximum likelihood (ML) estimators of the corresponding theoretical (population) rates \( (R_{ij}) \), and as such consistent estimators, it follows (from Slutsky's generalization of the weak law of large numbers; see Cramer, 1946, p. 235) that Expression 1' converges stochastically to

\[
\frac{\sum k_{ij} R_{ij}}{\sum k_{ij} R_{0j}}
\]

(1'')

This constant does not, in general, equal RR, unless the stratum-specific rates are uniform (homogeneous) \((i.e., R_{ij} = R_j, i = 1, 0)\); under this (rather unrealistic) condition Expression 1'' becomes the unbiased limit value:

\[
\hat{\text{RR}} = \frac{R_i \sum k_{ij}}{R_{0i} \sum k_{0j}} = \left( \frac{\text{RR}}{\sum S_{ij} / S_0} \right) \frac{\sum S_{ij} / S_0}{\sum S_{0j} / S_0} = \text{RR}.
\]

We may also note that the uniformity of the individual rate ratios \((\text{RR}_i)\) is not a sufficient guarantee for the pooled estimator to be consistent.

The above asymptotic result was first proven by Gart (1962) in a complete analogy for the odds ratio parametrization.

Recognizing the persistent problem with the pooled estimators of relative risk there arises the general question about the principles and procedures of combining strata so as to be able to construct optimal \((i.e., \text{both unbiased and efficient})\) estimators.

As a guideline, stratification is considered appropriate provided that the following three conditions obtain (Kleinbaum et al., 1982): the numbers of cases and noncases are sufficiently large in each stratum; a specified list of the stratification variates can be identified;

Public Health Reviews 1988: 16
for each variate a reasonable stratification scheme can be made (i.e.,
the number of strata is manageable without being overly wasteful of
information).

Given that the three conditions above hold, we shall proceed to
discuss how stratification analysis is carried out in statistical practice.

PREVAILING ANALYTIC PRACTICES FOR STRATIFIED
DATA

Principles, procedures, and products of data analysis: incoherent
approaches

The general goal of stratification analysis may be viewed as the
summarization of evidence in the data for the purpose of drawing an
inference. Specifically the concern here is with a uniform (over strata)
relation parameter \( (P) \), modeled either as rate difference (RD), rate
ratio, or odds ratio (OR). The usual aspects of data analysis with
these parameters are: (a) significance testing of the null hypothesis
\( (H_0: RD = 0, \ RR = 1, \ OR = 1) \); (b) point estimation of the
combined value of the parameter; and (c) computation of a confidence
interval for it.

The use of interval estimation over tests of significance should be
emphasized for two main reasons. First, statistical significance is
different from scientific (medical) importance and therefore estimation
of the magnitude of parameters is, in general, essential regardless of
whether significant \( p \)-value is obtained. As the main or only summary
of a statistical analysis, significance tests are less informative than
interval estimates and more liable to a moot interpretation. (Cf. Cox,
1977.) Second, as the confidence intervals are based on the same
function that is used to test postulated values of the parameter, it is,
of course, simple to deduce, from the relative location of the null
value of the parameter with respect to the interval, the extent the \( H_0 \) is
contradicted.

Prevailing analytic practices do not universally reflect the inherent
relation between estimation and testing. Moreover there is a tendency
to employ different principles for the derivation of these three types
of statistic. For example, with point estimates one is commonly
concerned with attainment of an unbiased expectation, whereas confidence intervals are designed, hypothetically in multiple repetitions of the study, to have a designated frequency of covering the theoretical value of the parameter. Theory also requires that, for example, the variance of the empirical value of the parameter be estimated acknowledging its dependence on the theoretical value of the parameter itself, irrespective of whether this is the null value of a value far in the region implied by the alternative (nonnull) hypothesis. Yet most procedures in current use, for instance in epidemiologic research, do not abide by this theory.

From this practice it follows that data-analytic results produced by methods commonly applied for the comparison of rates are incoherent. They can also be quite poor, even in the case of unstratified data (see Miettinen & Nurminen, 1985).

In the following sections the rationale of the main approaches currently in use are briefly commented on. The discussion is theoretical but nonmathematical.

Test statistics

Evidence from independent studies may be combined in several ways; the major methods — not necessarily having to do with occurrence-type data — have been reviewed by Rosenthal (1978). But different methods will produce inconsistent results when they are based on different functions of the data. According to the general theory of testing hypotheses, the construction of the best statistic should be made on account of the alternative hypothesis.

In cases when the parametrization of this alternative is a difficult problem to overcome, one may resort to the combining of p-values with the aid of Fisher's (1944) "omnibus" test procedure. But this rather ingenious method does not take into account the directions of the stratum-specific differences, and thus can mean loss of power. Nor does it appreciate the possible diversity of the independent studies (quality of information, attainment of comparability of the index and referent series, tenability of the underlying assumption of asymptotic chi square distribution, etc.). A final limitation of this procedure is that it is a technique of combining significance tests. The attitude governing today's statistical thinking is that the role of p-
value computation is overemphasized at the expense of effect estimation (Cox, 1977).

Although the topic of combining study results is a general one, in the next discussion we shall restrict the investigative context to the case of noneperimental occurrence research considered by Cochran (1954, p. 44):

"Suppose that we are comparing the frequencies of some occurrence in two independent samples, and the whole procedure is repeated a number of times under somewhat different conditions. The data then consists of a series of 2×2 tables, and the problem is to make a combined test of significance of the difference in occurrence rates in the two samples."

Based on a rather informal reasoning Cochran constructed a chi-square criterion that would be powerful if the alternative hypothesis implies a constant difference on a transformed (e.g., logit) scale for the compared rates. A modified version of this test is better known as the Mantel–Haenszel (1959) statistic.

The probability model under which Cochran derived his test statistic is J-fold product of pairs of independent binomial samples with the tables having marginal frequencies (i.e., total numbers of cases c_i and noncases S_i – c_i) free from any fixed conditions imposed on them. This model is applicable in cohort studies with proportion-type rates. The comparative parameter chosen by Cochran was a difference in rates. In the context of stratified data the null chi square statistic on one degree of freedom (d.f.) can be expressed as

\[
\chi^2_0 = \left( \frac{r_i - \hat{r}_i}{\hat{r}_i} \right)^2 / \left[ \sum W_i (1 - r_i) / (\Sigma W_j)^2 \right],
\]

(2)

where the mutually standardized rates \( \hat{r}_i = \sum_0 W_i r_i / \sum_0 W_i \), \( r_i = c_i / S_i \) and \( W_i \) are the Cochran null weights for the \( j \)th stratum, \( (1 / S_i + 1 / S_0)^{-1} \). Such weighting represents perfect proportionality to the amount of comparative information (i.e., the inverse of sampling variance) only when the alternative hypothesis is constant rate difference on logit scale (or, equivalently, constant odds ratio) over the strata (Birch, 1964; Radhakrishna, 1965). Yet by means of this assumption, Cochran could avoid the unstable estimation of the stratum-specific rate parameters (\( R_i \)) for the null chi-square with constant nonnull values for RD or RR as the alternative.
Apparently, Cochrane's aim was to construct a test that would be asymptotically efficient in detecting a common small nonzero value of the rate difference underlying all the studies (strata) to be combined. If this procedure is used only for the purpose of significance testing of $H_0$: $RD = 0$, then it is close to being most powerful among all unbiased tests. However, when stratification is done in the interest of validity (i.e., assuming a set of nonuniform $RD$'s), the estimation of the variance of the test statistic should involve the combined (nonnull) value of $RD$ instead of the null-hypothesis value of zero.

Mantel and Haenszel (1959) approached the same problem in another way. They assumed a conditional hypergeometric model with fixed marginal totals $(r_i = c_i/S_i)$. As a result their method employs actual population variances in place of the estimated variances used by Cochrane. Incidentally, the maximum likelihood (ML) estimates of the variances accepted by Cochrane are slightly biased and Mantel and Haenszel replaced them by the unbiased ones on multiplication by a factor of $S_i/(S_i - 1)$ in the denominator of the $\chi^2$ statistic (Expression 2). This modeling expanded the application of the Cochran–Mantel–Haenszel (hereafter CMH) procedure to case–control studies as well; OR is the “natural” parameter of the noncentral hypergeometric distribution. The general CMH principle has many desirable properties as noted above, but it may also be criticized in one important respect. The $\chi^2$ (Expression 2) is a null chi-square statistic and as such it may prove to be inaccurate when one departs far from the null state, as already alluded to above. Especially in occupational settings the exposures can be heavy and may cause high risks among workers in the extreme exposure categories. Thus, in these situations the actual (null or nonnull) value of the relation parameter should be used in order to arrive at accurate (variance) estimates.

Another reason for this inaccuracy stems from the fact that the CMH test assumes symmetry of the underlying sampling distribution (Gaussian) for the number of exposed cases ($c_{ij}$) in the individual strata. However, in epidemiologic practice the distribution of the $c_{ij}$ can be very skewed to the right (near Poisson). In these cases the consequence is that the $\chi^2$ exaggerates the observed level of significance; in other words, it gives overly significant $p$-values (see the section “Illustration” for an example).
The CMH test has been identified as a logit score statistic (Day & Byar, 1979). A major alternative procedure is the likelihood ratio test of \( H_0: \text{OR} = 1 \) under the hypothesis of “no second-order interaction” (Goodman, 1969), which is the same as common odds ratio in the \( J \) 2×2 tables (i.e., \( \text{OR}_1 = \text{OR} \)). However, an iterative algorithm (Bishop et al., 1969) is needed to find the expected cell frequencies when the condition \( \text{OR}_1 = \text{OR} \) is assumed.

A general way to approach the test problem is to consider an estimator of the relation parameter (possibly transformed), \( \hat{P} \), its deviance from the null value, \( \hat{P}_0 \), and expressed in units of the estimated null-variance of \( P \), \( \hat{V}_0(\hat{P}) \). The test statistic will then take the form of (see Kleinbaum et al., 1982)

\[
\chi^2_0 = (\hat{P} - \hat{P}_0) / \hat{V}_0(\hat{P}).
\]  

(3)

Particularly if the stratum-specific sample sizes are small, the general \( \chi^2_0 \) formulation (Expression 3) is statistically inferior to the CMH \( \chi^2_0 \) statistic (Expression 2). In addition, Expression 3 employs a wrong denominator (i.e., a variance estimate versus a variance parameter).

Point estimators

An overall (summary) estimate of a relation parameter (\( P \)) that combines information over all strata is usually feasible to compute only when the assumption of a uniform model (i.e., \( P_j = P \)) is tenable.

In principle there are two approaches to point estimation of overall relation, namely the methods of (a) maximum likelihood and (b) weighted averages. The various ML estimates (Norton, 1945; Birch, 1964; Gart, 1970) need an iterative (trial and correction) solution (Gart, 1971). However, using an approximation to likelihood equation of the log(OR), Birch (1964) derived a noniterative estimator for which it is not necessary to assume that the numbers of subjects \( S_j \) are large. But Birch’s estimator is guaranteed to be accurate only when the uniform OR is close to 1. The weighted average estimators can be divided into three classes (Kleinbaum et al., 1982): (i) precision-based estimators; (ii) Mantel–Haenszel estimators (Mantel &
Haenszel, 1959), and (iii) standardized estimators (Miettinen, 1972).

A general weighted average estimator (i) is expressible as (cf. Gart, 1962; Nurminen, 1981)

\[ \hat{P} = \frac{\sum W_i \hat{P}_i}{\sum W_i} \]  

(4)

In crude terms, weighting for precision means that each stratum of the data is weighted suitably according to the amount of information it contains (i.e., the inverse of the variance of \( \hat{P}_i \)).

The simple noniterative estimators, such as the difference of the weighted average of stratum-specific rates \( \hat{K}_D = r^* - r_0^* = \frac{\sum W_i \hat{K}_i}{\sum W_i} \), with \( r_i^* = \frac{\sum W_i r_i}{\sum W_i} (i = 1, 0) \) and the precision-maximizing null weights

\[ W_i = \left(1 / S_{ij} + 1 / S_0 \right)^{-1} \],

(5)

are meant to be used in instances when the uniformity hypothesis is reasonable, but it avoids the computational task needed to obtain an ML estimator. Analogously, a simple point estimator for \( \hat{K}_R \) is obtained by taking (see Greenland, 1982) \( \hat{K}_R = r_i^* / r_0^* = \frac{\sum U_i \hat{K}_i}{\sum U_i} \), with \( U_i = W_i / r_0 \).

Of the above two estimators (i and ii), the weights of the MH estimators are said to take into account both precision and importance, but the system lacks conceptual clarity (see Nurminen, 1981). The standardized estimators provide an unconfounded summary comparison with a known standard population. In particular, of the various summary odds ratios in use in case-referent studies only the MH estimator,

\[ \hat{\text{OR}}_{\text{MH}} = \frac{\sum W_{ij} (S_{ij} - c_{ij}) / S_j}{\sum W_{ij} (S_{ij} - c_{ij}) / S_j} \]

(6)

and the Miettinen standardized estimator,

\[ \hat{\text{OR}}_S = \frac{\sum W_{ij} (S_{ij} - c_{ij})}{\sum W_{ij} (S_{ij} - c_{ij})} \]

(7)
with the weights, \( W_s \), derived from the noncase subjects, do not assume uniformity of stratum-specific OR's. In the particular case of uniformity, whether or not we condition on the marginal totals, we get the same estimated variance for \( \hat{OR}_{MH} \) (Guilbaud, 1983), and \( \hat{OR}_{MH} \) is an almost fully efficient estimator of the common value. In the case in which \( OR_i = OR \), \( \hat{OR}_i \) is less efficient than \( \hat{OR}_{MH} \), especially in small samples, but, with large \( S_i \), the former estimates an epidemiologically more relevant population parameter (Greenland, 1982). On the other hand, the estimator \( \hat{OR}_{MH} \) has a connection to the CMH test so that \( \hat{OR}_{MH} = 1 \) only when \( \chi^2 = 0 \) (Expression 2), and thus unifies these two aspects of analysis.

Implicit in the original formulation of the Mantel–Haenszel estimator (Expression 6) was that some degree of non-uniformity is present. The estimator \( OR_{MH} \) also was “intended to be a weighted average of the separate table odds ratios rather than an estimate of the common odds ratio.” (Mantel, 1977, p. 189.) In spite of this, Tarone (1981) made explicit use of the null proposition of no relative effect (i.e., \( R_{ij} = R_{00} = R \)) to derive an analogue Mantel–Haenszel estimator of summary rate ratio (\( RR_{MH} \)) for cohort data. However, the assumption of an underlying uniform RR is not in keeping with the original intended application of the MH-type summaries. Moreover, the system of weights rests on the \( H_0: RR = 1 \), a condition so compelling that, if it were strictly true, it would make the \( RR_{MH} \) asymptotically fully efficient.

When stratification is needed only for the control of extraneous covariates, that is, for example, the variation in the \( RR \)'s materially represents sampling variation, we may assume that all strata provide information about the same underlying value of the RR parameter. If this were the case, the estimation of the summary ratio would be a matter of weighting by stratum-specific precision only, as pointed out by Mantel and Haenszel (1959, p. 735).

An exhaustive coverage of available methods for stratified analysis of relative measures is found in the excellent textbook by Kleinbaum et al. (1982).

**Interval estimators**

Just as in point estimation there are several alternative approaches to
interval estimation of comparative rate parameters. The two principle asymptotic methods rely on different kinds of approximations to the variance of the empirical measure; either (a) first-order Taylor series expansions or (b) test-based estimation (Miettinen, 1976a).

As for the latter approach (b), it yields interval estimates of the general form

$$\hat{p} \pm z_{\alpha/2} \hat{SE}$$

where $z_{\alpha}$ is the 100(1 - $\alpha$) centile of the Gaussian distribution and $\hat{SE}$ is a square root of some large-sample significance test (e.g., the CMH null chi-square statistic, Expression 2, bias-corrected). The procedure is easy to apply but, strictly speaking, it is valid only in the vicinity of the null-hypothesis value $P_0$, as it is derived assuming $P = P_0$. The method has the reasonable property that when the lower limit of a 100(1 - $\alpha$)% confidence interval equals the null value $P_0$, the lower-tail p-value is exactly $\alpha/2$. On the other hand the test-based principle has been strongly criticized (Halperin, 1977; Gart, 1979; Kleinbaum et al., 1982). The essence of the problem lies in that it is incorrect in the derivation of Expression 8 to equate two different asymptotic tests (a Wald-type statistic and, e.g., the CMH statistic) when considering a fixed alternative value of the relation parameter (Greenland, 1984). As it is thus valid to apply the test-based method only in the locality of $P_0$, it is prudent to use it generally as a rapid way to obtain rough interval estimates.

Regarding the use of a logarithmic-transformation along with a Taylor series approximation to the moments of the empirical statistic, this procedure, too, is problematic: in small samples the usual interval estimators (e.g., Woolf, 1955; Gart, 1962; Katz et al., 1978; Nurminen, 1981; see Kleinbaum et al., 1992) may be inaccurate or even fail in the presence of zero frequencies (see, e.g., Katz et al., 1978; Gart & Thomas, 1982; Nurminen, 1984). The fundamental reason for the inaccurate performance of these methods is that the estimation of the variance denominator of the empirical test statistic does not depend on the actual value of the relation parameter itself. An exception is the iterative method of interval estimation for OR of Cornfield (1956) and Gart (1970, 1971) who argued that it should be conditional on the total number of observed cases.
When the validity of the Gaussian distribution employed in the asymptotic methods becomes questionable there are two choices to correct the approximation: either to work out the conditional distribution based on the non-central hypergeometric distribution or to transform the test statistic. Regarding the former choice, a computing routine for the interval estimation of OR has been written by Thomas (1971), but there exist no statistics that are “M-ancillary” with respect to the RD parameter (Barndorff-Nielsen, 1976). As for the latter possibility, the logarithmic- and the angular-transformation are used to cope with the problems of skewness of the underlying distribution and nonconstancy of the variance estimate, respectively. In either case, however, the nonlinearity of the transformation may bring about intricacies of its own.

An attempt to tackle these difficulties is outlined in the following section.

**MODEL-BASED APPROACH**

The basic approach

A chi-square function-based method of analysis for rate measures was proposed recently by Miettinen and Nurminen (1985). The principle of the analysis, which by no means is new (cf. Wilson, 1927), is sketched as follows. Over various contemplated values of the model parameter $\theta$, the observed rates, $r_1$ and $r_0$, imply a chi-square function with one d.f. $\chi^2_1$ (asymptotic). The chi square is a single-valued function of the parameter $\theta$ for the given data ($c_i, s_i, i = 1, 2, \ldots, J$). Evaluated at the null value $\theta = \theta_0$, this function yields the null chi-square $\chi^2_{01}$, and this result corresponds to the $p$-value against $\theta_0$. The point estimate $\hat{\theta}$ is the parameter value for which $\chi^2_{01} = 0$. Finally, a $100(1 - \alpha)%$ two-sided confidence interval (equi-tailed), which is by far the most usual interval estimate in epidemiology, is demarcated by the two values of $\theta$ at which $\chi^2_{01}$ equals the $100 (1 - \alpha)$ centile of the chi-square distribution with one d.f. In particular, approximate 95% confidence limits are solved by setting $\chi^2_{01} = 3.84$. Thus the general chi-square function, $\chi^2_0$, provides all the elements that constitute the usual products of epidemiologic
data analysis. This means that fewer concepts are needed for inference.

The principal difference between the approach suggested by Miettinen and Nurminen (1985) and the prevailing approaches is in the way the variance denominator of the chi-square test statistic, on which the estimation procedure is based, is calculated. The usual methods employ separate ML estimators for the two sets of compared rates (R_{ij} and R_{ik}) without invoking any hypothesis about their relative magnitude. In comparison, in the new method the likelihood-based estimation is carried out conjointly for the rate parameters constrained by the actual value of the model parameter of interest, either RD, RR, or OR.

The advantages of the model-based approach are evident: (a) consistency of the different modes of analysis (significance testing, point and interval estimation); (b) uniformity of the chi-square function with respect to the three relation parameters (RD, RR, OR); (c) unnecessary or untenable assumptions are avoided (conditional on fixed margins, model of uniform risk, symmetry of distribution); (d) good overall performance, even in nonstandard situations (i.e., \( P \) in the nonnull region, \( S_{ij} \) small and variable); and (e) structured as a simple difference statistic: direct interpretation as well as easy to transform rates and to insert (e.g., bias) correction factors. On the other side there are the issues of: (f) need for an iterative estimation scheme; and (g) computation of the restricted ML estimates of the rates. Given that modern computing facilities are available, the latter points, (f) and (g), are, by and large, nonissues.

There are fairly appealing arguments for basing such a unified analysis on the likelihood for whatever the appropriate parameter is, the likelihood often being calculated conditional on the marginal constellations. In a great majority of cases, no doubt, likelihood ratio-based and efficient score-based methods and methods based on the proposed chi-square functions (employing a simple contrast as the numerator) give very similar answers; asymptotically the solutions are equivalent. In the score approach, for OR the conditional argumentation and the unconditional one are interchangeable; see Miettinen and Nurminen (1985) for details.

An advantage of the likelihood system is the mechanization it offers when various sets of data giving information on a common
parameter are to be combined for estimation. If \( L \) is the joint likelihood based on all the data and \( L_j \) for the \( j \)th stratum (part), then

\[
L = \prod_{j=1}^{J} L_j
\]

and, thus the total efficient score

\[
\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^{J} \frac{\partial \log L_j}{\partial \beta}.
\]

An element of the total information matrix is also obtained by simple addition of the corresponding pieces of information for each stratum (Rao, 1973, p. 374). Thus, in the stratified analysis the estimation based on the efficient score is a straightforward generalization of the unstratified procedure. The score statistic for OR yields exactly the same interval estimates as the iterative method of Cornfield–Gart, if the latter is supplemented with the \( S_j / (S_j - 1) \) bias correction factor and the \( 1/2 \) correction factor is deleted. In the case of RD and RR, however, the score-based method breaks down in the presence of zero margins in any stratum. For a detailed discussion of the likelihood-based approaches, see Miettinen and Nurminen (1985).

**Chi-square function**

In general terms, the weighted chi square function of the relation parameter (on one d.f.) suggested by Miettinen and Nurminen (1985) takes the form (cf. Expression 2).

\[
\chi^2 = \frac{\left( \sum W_i \delta_i / \sum W_i \right)^2}{\frac{\left( \sum W_i \delta_i / \sum W_i \right)^2}{\sum W_i \delta_i \hat{V}_i}}.
\]

with the difference (contrast) \( \delta_i = r_i - f_i(r_0) \); \( f_i \) stands for the generic function of the parameter \( P \). Specifically for rate difference \( (P = \text{RD}) \) the function is \( r_0 + \text{RD} \), in terms of + RD; in terms of rate ratio \( (P = \text{RR}) \) it is \( r_0 \) (RR), and for odds ratio \( (P = \text{OR}) \) it is \( r_0 \) (OR) / \( [1 + r_0 \text{ (OR - 1)}] \). The asymptotic variance estimator of the empirical difference \( \delta_i \) is

\[
\hat{V}_i = \left( \hat{V} \left( r_i \right) + \hat{V} f_i \left( r_0 \right) \right) S_i / \left( S_i - 1 \right),
\]

*Public Health Reviews 1988, 16*
with \( \tilde{V}(r_i) = \tilde{R}_i (1 - \tilde{R}_i) / S_i \) and the computing formulas for the restricted ML estimators (\( \tilde{R}_i \)'s) are given in the Appendix. The factor \( S_i / (S_i - 1) \) serves to provide for the unbiasedness of the variance estimator. The assumed unvarying weights (\( W_i \)'s) are chosen proportional, as closely as possible, to the amount of comparative information in the \( i \)-th stratum, that is, to the inverse of \( \tilde{V}_i \), obtained as the negative of the expected value of the second derivative of the log-likelihood function (Fisher information) evaluated at \( \tilde{R}_i \):

\[
W_i \propto \tilde{V}_i^{-1} \left[ \tilde{V}(r_i) + \tilde{V}(r_0) [f_r(\tilde{R}_0)]^2 \right],
\]

(11)

where \( f_r(\tilde{R}_0) = \partial f_r(\tilde{R}_0) / \partial \tilde{R}_0 \) is the partial derivative of the parameter function with \( P \) the relation parameter. In other words, this latter term reflects the extent to which the theoretical (reference) rate (\( \tilde{R}_0 \)) changes with the relation parameter. The particular forms of \( f_r(\tilde{R}_0) \) are: 1 in case of \( P = RD; RR \) for \( P = RR \); and \( OR / [1 + (OR - 1) \tilde{R}_0]^p \) for \( P = OR \). In Expression 11' the factor \( S_i / (S_i - 1) \), a constant over strata, is omitted as inconsequential.

The ideal weight (Expression 11) involves the unstable estimators of the unit variances. The problem can be avoided if it can be assumed that the \( \tilde{R}_i (1 - \tilde{R}_i) \) is constant over the strata. At any rate, this function changes relatively little when \( \tilde{R}_i \) lies anywhere between 30 and 70%. Thus the weights can be defined as (Miettinen & Nurminen, 1985)

\[
W_i^{-1} = 1 / S_i + \left[ f_r(\tilde{R}_i^*) \right] \left[ \tilde{R}_i^* (1 - \tilde{R}_i^*) / \tilde{R}_i (1 - \tilde{R}_i) \right] / S_i
\]

(11')

with \( \tilde{R}_i^* = \sum W_i \tilde{R}_i / \sum W_i, \) \( i = 1, 0. \) However, the \( W_i \) need to be solved iteratively. In Expression 11' the factor \( \tilde{R}_i^* (1 - \tilde{R}_i^*) \), a constant over strata, is omitted as inconsequential.

For the pooled analysis of rates the chi-square function in Expression 9 simplifies to \( \chi^2_P = d^2 / \tilde{V} \) with \( d = r_i - \hat{f}_P(n_i) \) and \( \tilde{V} \) as in Expression 10 but the subscript \( j \) deleted.

The results of comprehensive simulation experiments for unstratified (Nurminen, 1984; Miettinen & Nurminen, 1983) confirmed the better overall accuracy of the \( \chi^2_P \) function-based interval estimates over their usual counterparts.
On the null hypothesis the contrast-based $\chi^2_0$ function (Expression 9) for the different relation parameters reduce to the same CMH null chi square (Expression 2, bias corrected). Insofar as the contemplated value of the relation parameter is not too far away from the null state, it is worthwhile to transform the empirical rates in order to make the distribution of $\chi^2_0$ approach the assumed Gaussian form more closely and rapidly. However, the normalizing transformations proposed by Borges (see Johnson & Kotz, 1969, p. 65) is not easily practicable. Fortunately, variance-stabilizing transformations commonly normalize as a by-product, even though they do not produce the optimum normalization (Kendall & Stuart, 1966). To this end we shall make use of the arcsine transformation for a binomial rate or $t(r) = \arcsin(r)^{1/2}$, and thus get (Miettinen, 1983; Nurminen, 1984):

$$\chi^2_0 = \left[ t \left( t^*_1 \right) - (r^*_1) \right] / \left( \sum_i W_i c_j \left( t \right) \right) \left[ S_i / \left( S_i - 1 \right) \right] / \left( \sum_i W_i \right)^2 \left[ 1 / \left( 4 r^*_1 \left( t - r^*_1 \right) \right) \right],$$

where $r^*_1 = \sum_i W_i c_j / \sum_i W_i$.

Generalizations to the nonnull situation are manageable but involve complexities which are beyond the scope and level of this paper (see Nurminen, 1984).

**ILLUSTRATION**

To illustrate the impact of the alternative procedures consider the cumulative incidence data in Table 1 analyzed previously by Miettinen (1983) and Nurminen (1984).

Because the empirical rate difference is 7.9% among males and 24.4% among females the concern here is with stratified analysis in the interest of comparability rather than estimation of a common value of the theoretical rate difference. The numbers of subjects in each stratum ($S_i$) are apparently large enough for the Gaussian approximation to be adequate. Moreover, as $S_0 \gg S_1$ for each stratum, $W_0$ from Expression 11 is approximately equal to $S_0$, a
Table 1

Incidence of drug-attributed rash in relation to allopurinol exposure among recipients of ampicillin, by gender

<table>
<thead>
<tr>
<th>Stratum of rate</th>
<th>Component</th>
<th>Exposure to allopurinol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (j=1)</td>
<td>Cases</td>
<td>5(c1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36(c0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41(c)</td>
</tr>
<tr>
<td></td>
<td>Noncases</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>645</td>
</tr>
<tr>
<td></td>
<td></td>
<td>678</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>38(S1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>681(S0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>719(S)</td>
</tr>
<tr>
<td>Female (j=2)</td>
<td>Cases</td>
<td>10(c1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38(c0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>68(c)</td>
</tr>
<tr>
<td></td>
<td>Noncases</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>518</td>
</tr>
<tr>
<td></td>
<td></td>
<td>537</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>29(S1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>576(S0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>605(S)</td>
</tr>
</tbody>
</table>


constant, regardless of the value of the rate difference parameter. For instance, with RD = 0 one obtains $W_1 = 35.9$ and $W_2 = 27.6$, while with RD = 0.3, $W_1 = 37.4$ and $W_2 = 28.5$. If a common underlying value for rate difference is assumed, it is estimated to be 0.15. The corresponding unconstrained ML estimate (due to Miettinen, see Rothman & Boice, 1978) is 0.14. Symmetric about this point are the 95% test-based limits (Expression 8) for the parameter: 0.08 and 0.22. These limits are accurate in the locality of the null hypothesis (RD = 0) but are known to produce quite inaccurate (too narrow) interval estimates when the actual value of the parameter departs far from the null state. Here the corresponding limits based on chi-square function of rate difference defined by Expressions 9 through 11 are wider, 0.07 and 0.26. Especially the upper limit is markedly more distal than the respective test-based value (see Fig. 1).

In terms of rate ratio the gender-specific point estimates are more homogeneous: 2.5 and 3.4 for males and females, respectively; the combined estimate is computed to be 3.1. The 95% confidence limits for the common value of RR are 1.76 and 4.70.

Since for both genders the null distribution of the number of exposed cases is asymmetric (i.e., both of the marginal incidence...
proportions are far from the value of 1/2) and the realizations (r_{ij}) are, typically, on the "long tail," the $\chi^2_{RD}$, which assumes symmetry of distribution, gives exaggerated values without a transformation of the rates. For instance, against RD = 0 the realization of the CMH statistic (Expression 2) is 19.5, a presumably excessive value as against that obtained for the angular $\chi^2_{A}$ (Expression 12), viz. 12.2 (see Fig. 1).

*Public Health Reviews* 1988; 16
As a final remark it should be appreciated that especially in occupational health settings, great accuracy in summarizing (pooling) methods is not of vital importance in view of the often large nonstatistical sources of differences among studies (Greenland, 1987). Thus, the analysis of pooled data should proceed from estimation of overall measures or testing for global null hypothesis of no relation at all to analysis for modification of an effect.

REFERENCES


Public Health Reviews 1988; 16


APPENDIX
Restricted maximum likelihood estimators of the rate parameters (unstratified data)

For RD, \( \tilde{R}_0 = 2 \frac{p \cos (\alpha)}{L_3} - \frac{L_2}{L_3} (3 L_3) \),

where \( \alpha = \frac{1}{3} \left[ \pi + \cos^{-1} \left( \frac{q}{p^3} \right) \right], \)

\( p = \pm \left[ \frac{L_1^2}{L_3} - \frac{L_1}{L_3} (3 L_3) \right]^{1/2}, \)

\( q = \frac{L_2^2}{L_3} - \frac{L_2}{L_3} (6 L_3^2 + L_2) / (2 L_3), \)

with the sign of \( p \) chosen so that the signs of \( p \) and \( q \) coincide, and the coefficients of the likelihood equation

\[ \sum_{k=0}^{3} L_k \tilde{R}_k = 0 \]

are:

\( L_3 = S, \)
\( L_2 = (S_1 + 2 S_0) (RD) - S - c, \)
\( L_1 = (S_0 (RD) - S - 2 c_0) (RD) + c, \)
\( L_0 = (RD) (1 - RD) c_0. \)

For RR, \( \tilde{R}_0 \) = \( \frac{B - (B^2 - 4 AC)^{1/2}}{(2 A)} \),

with \( A = S (RR), B = S_1 (RR) + c_1 + S_0 + c_0 (RR), C = c, \)

For OR, \( \tilde{R}_0 \) = \( \frac{-B + (B^2 - 4 AC)^{1/2}}{(2 A)} \),

with \( A = S_0 (OR - 1), B = S_1 (OR) + c_0 (OR - 1), C = -c. \)

In any case, \( \tilde{R}_1 = f (\tilde{R}_0). \)

Public Health Reviews 1988; 16