Optimal Weighted Estimation Versus Cochran-Mantel-Haenszel

Lisa Hermans[1*] **Geert Molenberghs**[1,2] **Geert Verbeke**[2,1]
Michael G. Kenward[3] **Pavlos Mamouris**[4] **Bert Vaes**[4]

[1] *I-BioStat, Data Science Institute, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

[3] *Luton, United Kingdom*

[4] *Academisch Centrum voor Huisartsgeneeskunde, KU Leuven, B-3000 Leuven, Belgium*

∗ *Corresponding author: Martelarenlaan 42 B-3500 Hasselt; lisa.hermans@uhasselt.be*

Abstract

The purpose of this paper is to contrast the Mantel-Haenszel estimator with an optimal estimator to better understand its specific nature, as well as some unique and interesting properties of the data setting for which it was developed.

It is emphasized here that the Mantel-Haenszel estimator does not follow from optimality considerations, but nevertheless has properties similar to and often better than the optimal estimator, whose existence we demonstrate in spite of the absence of completeness. It is shown, via simulations and data analysis, that the optimal estimator outperforms the Mantel-Haenszel estimator only in certain settings with huge sample sizes.

Some Keywords: Pseudo-Likelihood; Split sampling; Unequal strata sizes.

1 Introduction

Categorical variables take on an a priori fixed and finite number of possible values, in particular, two values for binary data. Investigators often want to examine associations between categorical variables, for example, in a case-control study. In the past, the need grew to analyse the relation between binary variables incorporating the classification due to other relevant confounders. Mantel

& Haenszel (1959) published their view on the analysis of case-control studies more than half a century ago. Their methodology has become ubiquitous in epidemiology and beyond.

The Mantel-Haenszel estimator (Mantel & Haenszel, 1959, MH) can be used in various contexts. It serves as the estimator for the odds ratio in a series of 2×2 tables or in matched designs. In general, a setting with stratified 2×2 tables already exists when subpopulations within the overall population vary. The resultant strata need not be of equal size. The simplest form of matching is 1:1, i.e., one control per case. This is effective when both are sufficiently prevalent in the population. Often, cases are more scarce than controls, e.g., with rare diseases. It then makes sense to select more controls per case. When a case is individually matched to a set of controls, having similar values for some confounding variables, the most extreme form of stratified design is created. Each case and corresponding control(s) can be seen as one stratum.

The MH is not fully parametric and serves to investigate the independence in a $2 \times 2 \times N$ contingency table (Agresti, 2002, pp. 231). Typically strata are of varying sizes, naturally necessitating the use of weights. Mantel & Haenszel (1959) proposed several weighting schemes to estimate a common odds ratio. It is important to realize that these weighting schemes, even though they have been of great use for decades, do not follow from formal optimality criteria. Interestingly, no expression for the variance was available until the work of Robins *et al.* (1986b). With time, others investigated the estimator and many extensions have been developed. Kuritz *et al.* (1988) reviewed the MH and its variance formulas. At first, Hauck (1979) proposed an estimator of the variance using a product binomial model, appropriate for large stratum samples. Woolf (1955) used a logarithmic transformation of the odds ratio estimator, which makes the sampling variance simple and easy to use as weights. From another point of view, Flanders (1985) proposed a variance estimator based on a series of Monte Carlo experiments, leading to more accurate confidence intervals. Robins *et al.* (1986a,b) proposed a new robust variance estimator based on the unconditional distribution of the data. These last two are very similar and even identical for matched designs. Either is applicable in both sparse data and large-strata limiting models and are easily computed. None of these estimators had been formally shown to be "best," but the latter two are preferred. Our goal is to place the MH against the background of estimators that do follow from optimality considerations.

Hermans *et al.* (2018a,b) considered estimation in hierarchical data settings with unequal cluster sizes with compound-symmetry (CS) and AR(1) covariance structures. When cluster sizes are not of the same size, complete sufficient statistics typically fail to exist (Casella and Berger, 2001, pp. 285–286) and maximum likelihood estimators (MLEs) do not have closed-form solutions. This has

Table 1: Contingency table for stratum i $(i = 1, \ldots, N)$

	Exposure +	**Exposure -**	Total
Case +	a_i	b_i	n_{1i}
Control -	c_i	d_i	n_{2i}
Total	m_{1i}	m_{2i}	n_i

important implications for the analysis of very large sets of data for which efficient computation is critical. We are therefore interested in performance comparisons under a wide range of settings, from small and moderate sample sizes to big data settings. Molenberghs *et al.* (2011) proposed a pseudo-likelihood split-sample based approach, where the original sample is divided into subsamples within each of which MLEs are consistent estimators. Subsample-specific results are then combined using appropriate optimal weights. This idea is closely related to the MH for grouped data, where the group-specific odds ratios are combined using weights. Even though the idea is similar, there are computational differences. For this reason the determination of the optimal weights could be extended to these grouped data settings. All statistical procedures mentioned in this paper are appropriate for data settings with binary responses, involving stratification, grouping, or matching based on confounding variables.

Throughout the paper, we assume that the strata are mutually independent, in line with the original formulation of the MH and its subsequent use (Agresti, 2002), and as such ensuring that our focus is on a comparison with that original form.

The remainder of the paper is organised as follows. In Section 2, the general methodology of the MH is given, together with the notation that will be used throughout the paper. In Section 3, pseudo-likelihood methodology is applied and weighting schemes are explored. In Section 4, a simulation study is described to compare the performance of the MH with one following from optimal weighting considerations. An example is discussed in Section 5. The discussion and recommendations for practice are offered in Section 6.

2 Mantel-Haenszel Estimator

The Mantel-Haenszel estimator (Mantel & Haenszel, 1959) is a useful and convenient estimator for obtaining a common odds ratio, when there are one or more confounders. When the sample is stratified by one or more variables, the subgroup-specific odds ratios are combined into a weighted average. In the case of several, N say, 2×2 tables or strata, the i^{th} $(i = 1, \ldots, N)$ stratum takes the form as presented in Table 1. The overall sample size is defined as $n = \sum_{i=1}^{N} n_i$.

As pointed out earlier, Mantel & Haenszel (1959) suggested various weighting schemes, but the most widely accepted estimator of the common odds ratio is:

$$\tilde{\psi}_{MH} = \frac{\sum_{i=1}^{N} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{N} \frac{b_i c_i}{n_i}} = \frac{\sum_{i=1}^{N} w_i \frac{a_i d_i}{b_i c_i}}{\sum_{i=1}^{N} w_i}, \tag{1}$$

with $w_i = \frac{b_i c_i}{n_i}$.

The strata do not need to be of the same size and even if some cell counts are small or even zero, the estimator remains well-defined, an important asset. Also, when $b_i c_i$ equals zero a stratum is omitted in the calculation of the common odds ratio as the weight becomes zero as well. This estimator is very practical to use.

When originally formulated, no variance formula was available. Over time, proposals were developed and nowadays the variance formula of Robins *et al.* (1986b) is commonly used, particularly in statistical software. It will be used later in this paper and takes the following form:

$$v_R = \mathsf{var}(\log \hat{\psi}_{MH}) = \frac{\sum_{i=1}^{N} \frac{(a_i+d_i)a_i d_i}{n_i^2}}{2\left(\sum_{i=1}^{N} \frac{a_i d_i}{n_i}\right)^2} + \frac{\sum_{i=1}^{N} \frac{(a_i+d_i)b_i c_i + (b_i+c_i)a_i d_i}{n_i^2}}{2\left(\sum_{i=1}^{N} \frac{a_i d_i}{n_i}\right)\left(\sum_{i=1}^{N} \frac{b_i c_i}{n_i}\right)} + \frac{\sum_{i=1}^{N} \frac{(b_i+c_i)b_i c_i}{n_i^2}}{2\left(\sum_{i=1}^{N} \frac{b_i c_i}{n_i}\right)^2}. \tag{2}$$

3 Optimal Weighted Estimation

Breslow and Day (1980) examined likelihood estimation to obtain an estimator for the common odds ratios. However, they showed that there is no closed-form solution, except for the case where there are only two strata. Numerical estimation techniques are necessary in pursuing an estimate. In comparison with this, the MH is much simpler to use. Hermans *et al.* (2018a,b) use weighted estimation for data settings with unequal cluster sizes. Based on the pseudo-likelihood split-sample approach of Molenberghs *et al.* (2011), the sample is divided into subsamples, containing clusters of equal size. For each subsample, maximum likelihood estimators were calculated and subsample-specific results then combined using weights. The entire argument will not be reproduced, but these same ideas can be used for the data settings discussed in this paper. The subsamples considered here are naturally the various strata in the sample. The subsample or stratum-specific estimator is the odds ratio, ψ, calculated as:

$$\psi_i = \frac{a_i \cdot d_i}{b_i \cdot c_i}. \tag{3}$$

These can be combined into a common odds ratio using weights α_i:

$$\tilde{\psi} = \sum_{i=1}^{N} \alpha_i \psi_i, \tag{4}$$

with $\sum_{i=1}^{N} \alpha_i = 1$. It is natural and well known that optimal weights are inversely proportional to a measure of variance (see the Appendix for a brief sketch of the argument). We will now investigate this further, against the background of the lack of complete sufficient statistics.

3.1 (In)Complete Sufficient Statistics

Completeness means that any measurable function of a sufficient statistic, which has zero expectation for every value of the parameter indexing the parametric model class, is the zero function almost everywhere. The relevance of incompleteness can be found in two important theorems, the Lehman-Scheffé theorem (Casella and Berger, 2001) and Basu's theorem (Basu, 1955). In particular, the first one leads to mean-unbiased estimators when the conditions are fulfilled.

Consider the weighted odds ratio estimator as in (4), and assume $E[\psi_i] = \psi$, then

$$E[\tilde{\psi}] = \sum_{i=1}^{N} \alpha_i E[\psi_i] = \psi \sum_{i=1}^{N} \alpha_i = \psi.$$

Suppose that there is a non-zero function $g((\psi_i)_i) = \sum_i \beta_i \psi_i$, such that $E[g((\psi_i)_i)] = \sum_i \beta_i \psi = \psi \sum_i \beta_i = 0$. This is satisfied for all β_i's where $\sum_i \beta_i = 0$. By this counterexample, incompleteness holds. As a consequence, it is *a priori* not guaranteed that there is a uniformly optimal estimator. However, it should be noted that the existence of a uniform optimum, while not guaranteed by the theorem, is not necessarily excluded.

3.2 Optimal Weights

To obtain (potentially local) optimal weights, we seek to minimize the variance, $\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \alpha_i^2 \text{var}(\psi_i)$, under the constraint that $\sum_{i=1}^{N} \alpha_i = 1$ using the method of Lagrange multipliers. The calculations, which are standard and applicable in a wide variety of settings, are briefly reviewed in Appendix A. The weights are:

$$\alpha_i = \frac{v_i^{-1}}{\sum_j v_j^{-1}}. \tag{5}$$

In the next step, the variance of a stratum-specific odds ratio will be expressed explicitly. We here assume, in line with what was stated in the Introduction, that strata are mutually independent. When taking the natural logarithm of the odds ratio, the variance here equals $\text{var}(\log(\psi_i)) = \frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}$. By using the delta method we find $\text{var}(\log(\psi_i)) \cong \frac{1}{\psi_i^2} \cdot \text{var}(\psi_i)$ and now

$$\begin{aligned} \text{var}(\psi_i) &= \psi_i^2 \text{var}(\log(\psi_i)) \\ &= \psi_i^2 \left(\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}\right) \\ &\overset{\text{pop. value}}{\cong} \frac{\psi^2 q}{n_i}, \end{aligned} \tag{6}$$

with $q = \frac{1}{p_{11}}+\frac{1}{p_{10}}+\frac{1}{p_{01}}+\frac{1}{p_{00}}$, and p_{11}, p_{10}, p_{01} and p_{00} the 2×2 cell probabilities. Here, evidently, it is assumed, in line with the original MH, that all strata are generated from the same underlying contingency table, of which then the probabilities and functions thereof are constant, hence also $\psi_i = \psi$. This situation arises when an unstratified analysis would lead to a confounded relationship between row and column classification. Thus, we assume that the stratificiation undertaken is sufficient to control for the confounding that might exist at population level, and that there is no effect modification. Given this setting, the stratum-specific variance and its inverse equal:

$$\text{var}(\psi_i) \cong \frac{\psi^2 q}{n_i} = v_i \Rightarrow v_i^{-1} = \frac{n_i}{\psi^2 q}, \tag{7}$$

resulting in the following formula for the weights:

$$\alpha_i = \frac{\frac{n_i}{\psi^2 q}}{\sum_j \frac{n_j}{\psi^2 q}} = \frac{n_i}{n}, \tag{8}$$

with $n = \sum_{i=1}^N n_i$. If all $(n_i)_i$ would be fixed by design, this is a uniform minimal solution, in spite of incompleteness. Using the above expressions, it follows that the uniformly optimal weighted estimator satisfies:

$$\tilde{\psi} = \sum_{i=1}^N \frac{n_i}{n} \psi_i. \tag{9}$$

Equation (19) yields an expression for the overall variance of this common odds ratio:

$$\text{var}(\tilde{\psi}) = \frac{\sum_{i=1}^N v_i^{-2} v_i}{(\sum_{j=1}^N v_j^{-1})^2} = \frac{1}{\sum_{i=1}^N v_i^{-1}}. \tag{10}$$

This scheme is more principled and enjoys minimum variance properties, in comparison to MH. However, the estimator is undefined when there is at least one zero in one of the contingency tables. It suggests that MH may well be superior in small samples. Also, because MH does not take the form of a conventional weighted estimator, with weights differing from the optimal ones, also the behavior in large to very large samples should be investigated.

We can also use the observed rather than expected variances. Then, an alternative sub-optimal form emerges:

$$\tilde{\tilde{\psi}} = \sum_{i=1}^{N} \alpha_i \psi_i = \frac{\sum_{i=1}^{M} v_i^{-1}\psi_i}{\sum_{i=1}^{M} v_i^{-1}}. \tag{11}$$

Now,

$$\begin{aligned} v_i &= \psi_i^2\left(\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}\right) \qquad (12)\\ v_i^{-1} &= \psi_i^{-2}h_i, \end{aligned}$$

with

$$h_i = \left(\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}\right)^{-1} = \frac{a_ib_ic_id_i}{e_i}, \tag{13}$$

with $e_i = a_ib_ic_i + a_ib_id_i + a_ic_id_i + b_ic_id_i$. This leads to:

$$\tilde{\tilde{\psi}} = \frac{\sum_i \frac{(c_ib_i)^2}{e_i}}{\sum_i \frac{1}{\psi_i}\frac{(c_ib_i)^2}{e_i}}. \tag{14}$$

Yet another sub-optimal estimator, based on MH ideas, would be:

$$\tilde{\tilde{\psi}}' = \frac{\sum_i a_id_i\frac{(c_ib_i)^2}{e_i}}{\sum_i c_ib_i\frac{(c_ib_i)^2}{e_i}}. \tag{15}$$

We choose not to present variance formulae for these alternative estimators (14) and (15) because, as we see below, their relative performance is so poor as to obviate their use in practice. This underscores that the choice of weights should be done very judiciously and consequently that a given choice should be investigated carefully, in theoretical terms and/or using simulations. The demonstration of poor perfomance with slightly deviating weights can still motivate further research.

4 Simulation Study

In the following simulation study, carried out in R, the performance of the derived estimators is compared with the conventional MH.

To start, a sample with five strata and stratum sizes 900, 500, 200, 600, and 300 is considered. For each stratum, random numbers following a uniform distribution are generated, as many as the

Table 2: The underlying 2×2 probability table used in the simulation study

p_{00} =0.2	p_{01} =0.1
p_{10} =0.4	p_{11} =0.3

stratum size requires. These numbers are classified according to the cumulative probabilities (0, p_{00} =0.2, $p_{00} + p_{01}$ =0.3, $p_{00} + p_{01} + p_{10}$ =0.7, $p_{00} + p_{01} + p_{10} + p_{11}$ =1) from the underlying contingency table (Table 2) to then yield the frequencies for a new stratum-specific 2×2 table. For example, the sampled number 0.4 will contribute to the count in cell (1,0) as it lies between 0.3 and 0.7. This results in five new 2×2 tables. To these tables, the MH and estimators (9), (14), and (15) are applied. In this scenario there were 5 strata with mean stratum size 500. In further samples, both the number of strata and mean stratum size are multiplied by a factor of ten. All values for numbers of strata (5; 50; 500; 5,000; and 50,000) and mean stratum sizes (500; 5,000; 50,000; 500,000; and 5,000,000) are combined into 25 scenarios. Recall that the very large sizes are considered to examine performance in big-data settings. Each scenario was sampled 500 times, leading to 500 estimates of the common odds ratio for each estimator. By doing this we can explore the estimators' performance under varying designs. As an aside, when the mean stratum size and number of strata get large, there simulation is time-consuming. As the performance of the estimators is already proven in the middle of the table, some of the bottom cells in the upcoming tables are left blank.

4.1 Relative Efficiency

As an important evaluation criterion, we consider the relative efficiency of the proposed estimators to the MH. First, we calculate the simulation-based variances of our estimators over the sample of 500 odds ratios. Second, the model-based variances are used. For the Mantel-Haenszel estimator the variance formula (2) is used, and (10) is used as our estimator. This choice is motivated by the fact that we aim to compare the true variances of these estimators; simulation-based variances are a fair way to do so, especially if sample sizes and numbers of strata are sufficiently large.

The results based on the simulation-based variances are presented in Tables 3–5. The results based on the model-based variances are reproduced in Table 6. These are for estimator (9), given the relatively poor performance of estimators (14) and (15). The relative efficiencies are presented as percentages.

Considering Tables 3–5, the MH is slightly more efficient for large samples but not for huge samples. Only at the right bottom corner of the table the relative efficiency nears 100%. For

Table 3: Simulation: Relative efficiency of estimator (9) w.r.t. Mantel-Haenzel estimator. (Simulation based.)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	90.51	98.62	100.12	100.02	100.00
50	90.42	98.71	99.52	100.03	99.96
500	89.61	99.12	99.51	99.96	100.02
5 000	90.93	99.19	100.17	99.92	
50 000	92.41	99.38	99.79		

Table 4: Simulation: Relative efficiency of estimator (14) w.r.t. Mantel-Haenszel estimator. (Simulation based.)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	97.65	101.44	99.57	99.92	99.99
50	94.08	100.35	100.86	99.86	100.07
500	90.30	99.17	100.77	100.06	100.01
5 000	90.14	98.97	99.25	100.20	
50 000	90.95	98.44	99.94		

Table 5: Simulation: Relative efficiency of estimator (15) w.r.t. Mantel-Haenszel estimator. (Simulation based.)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	69.65	75.31	65.65	72.73	73.28
50	66.07	71.80	64.36	68.85	72.04
500	67.04	73.18	72.28	64.97	69.57
5 000	70.98	66.53	70.02	66.33	
50 000	71.02	68.67	67.58		

Table 6: Simulation: Relative efficiency of (9) w.r.t. Mantel-Haenszel estimator. (Model based)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	106.58	100.65	100.06	100.01	100.00
50	108.50	100.80	100.08	100.01	100.00
500	108.63	100.83	100.08	100.01	100.00
5 000	107.66	100.83	100.08	100.01	
50 000	107.66	100.83	100.08		

extremely large overall sample sizes, the alternative estimators (9) and (14) are performing equally well as the MH. The opposite is found in Table 6. In the left top corner, relative efficiency exceeds 100%. The variance estimator (10) turns out to be smaller, and hence more precise, than the asymptotic variance of Robins *et al.* (1986b); at the same time it is easier to compute. Pattanayak *et al.* (2012) also showed in their simulation study that the variance estimator of Robins *et al.* (1986b) is conservative, producing unnecessarily wide confidence intervals. This simulation study confirms this. However, when the mean stratum sizes become very large, this issue goes away.

Turning to Table 5, estimator (15) is an ad hoc modification of (14) by using, at face value, MH ideas. However, while maybe a natural estimator to consider, the simulation study gives evidence that its performance is poor, evidently without given a precise mathematical explanation, other than that it is not following optimality arguments.

Note that the relative efficiency of (9) is different than that of (14). We should note, though, that both efficiencies are relative to the MH estimator. Neither MH nor (14) are optimal by any formal criterion, whereas (9) is. It is therefore to be expected that the relative efficiency of (9) is very good especially for large to huge sample sizes.

4.2 Coverage Probabilities, Bias, and Mean Squared Error

Next, the coverage probabilities, bias, and mean squared error (MSE) of the estimators are examined. Table 7 shows the coverage probabilities for estimator (9), where the 95% confidence interval is calculated using the proposed variance estimator (10) and normal quantiles. The use of the normal quantiles is supported by the Gaussian shape of the density plots in Figure 1. For comparison, Tables 9 and 10 give the coverage probabilities for the MH. Where Table 10, like Table 7, displays simulation-based coverage, in Table 9 the coverage is based on model-based confidence intervals; these are calculated for the log odds ratio first and then backtransformed to the original scale. These tables suggest that the proposed estimator is performing badly in the left bottom corner of the tables, as the coverage probabilities become small and even zero. The same phenomena can be observed in Tables 8 and 11 where the confidence intervals are calculated with the sample variance. The latter is done to make sure that there is no over-coverage, which does not seem to be the case. Tables 12–15 show, respectively, the bias and MSE of the alternative estimator and the MH. Here, the bias and MSE are larger in the same part of the tables.

Figure 1 supports earlier conclusions. In general, the plots reflect the poor behaviour of estimators (14) and (15). In relationship to Tables 7–15, the density plots in the left bottom corner are getting

Table 7: Simulation: Coverage Probabilities for estimator (9)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.940	0.950	0.950	0.930	0.950
50	*0.816*	0.932	0.946	0.936	0.942
500	*0.186*	*0.846*	0.940	0.964	0.932
5 000	*0.000*	*0.204*	*0.868*	0.948	
50 000	*0.000*	*0.000*	*0.204*		

Table 8: Simulation: Coverage Probabilities for estimator (9) with sample variance

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.938	0.940	0.946	0.938	0.956
50	*0.862*	0.946	0.944	0.940	0.954
500	*0.198*	*0.848*	0.944	0.956	0.930
5 000	*0.000*	*0.212*	*0.832*	0.948	
50 000	*0.000*	*0.000*	*0.200*		

worse.

It is hard to say exactly why. We do note that (9) is optimal among the weighted estimators, that (14) and (15) are not, and that MH does not even belong to the class of weighted estimators. Therefore, we cannot easily explain the relative behaviour of (9), (14), and (15) relative to MH. The only evident statement is that (9) is supposed to outperform the other two weight-based estimators.

5 Analysis of Case Study

Intego is a Belgian general practice-based morbidity registration network at the Department of General Practice of the University of Leuven, Belgium. They built a large database as a result of continual recording of data in general practices since 1994. It holds over 4 million diagnoses, 44 million laboratory results and 17 million medication prescriptions and 700,000 vaccination data.

Table 9: Simulation: Coverage Probabilities for Mantel-Haenszel estimator (log oddsratio)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.962	0.950	0.950	0.930	0.950
50	0.948	0.942	0.946	0.936	0.942
500	0.960	0.958	0.944	0.960	0.932
5 000	0.954	0.954	0.966	0.942	
50 000	0.954	0.940	0.958		

Table 10: Simulation: Coverage Probabilities for Mantel-Haenszel estimator

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.952	0.952	0.952	0.930	0.950
50	0.958	0.940	0.946	0.938	0.942
500	0.960	0.958	0.944	0.960	0.932
5 000	0.954	0.954	0.966	0.942	
50 000	0.954	0.940	0.958		

Table 11: Simulation: Coverage Probabilities for Mantel-Haenszel estimator with sample variance

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.954	0.948	0.950	0.938	0.956
50	0.956	0.954	0.946	0.942	0.958
500	0.956	0.956	0.948	0.956	0.932
5 000	0.948	0.954	0.952	0.942	
50 000	0.956	0.940	0.954		

Table 12: Simulation: Bias for estimator (9)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.045 487	0.006 383	-0.000 194	-0.000 174	0.000 067
50	*0.039 827*	0.004 114	0.000 356	-0.000 001	0.000 018
500	*0.040 240*	*0.003 951*	0.000 404	0.000 019	9.580e-07
5 000	*0.039 646*	*0.003 817*	*0.000 400*	0.000 025	
50 000	*0.039 552*	*0.003 840*	*0.000 386*		

Table 13: Simulation: MSE for estimator (9)

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.022 471	0.001 935	0.000 190	0.000 020	0.000 002
50	*0.003 755*	0.000 225	0.000 019	0.000 002	2.016e-07
500	*0.001 819*	*0.000 034*	0.000 002	1.762-e07	1.851e-08
5 000	*0.001 591*	*0.000 016*	*3.248e-07*	1.939e-08	
50 000	*0.001 566*	*0.000 015*	*1.674e-07*		

Table 14: Simulation: Bias for Mantel-Haenszel estimator

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.013 178	0.003 413	-0.000 491	-0.000 204	0.000 064
50	0.001 020	0.000 444	-0.000 012	-0.000 038	0.000 015
500	0.000 849	0.000 181	0.000 028	-0.000 018	-0.000 003
5 000	0.000 180	0.000 053	0.000 025	-0.000 013	
50 000	0.000 104	0.000 072	0.000 010		

Table 15: Simulation: MSE for Mantel-Haenszel estimator

Number of strata	Mean stratum size 500	5 000	50 000	500 000	5 000 000
5	0.018 639	0.001 880	0.000 191	0.000 020	0.000 002
50	0.001 962	0.000 205	0.000 018	0.000 002	2.014e-07
500	0.000 180	0.000 019	0.000 002	1.762e-07	1.852-e08
5 000	0.000 017	0.000 002	1.657e-07	1.894-e08	
50 000	0.000 002	1.929e-07	1.864e-08		

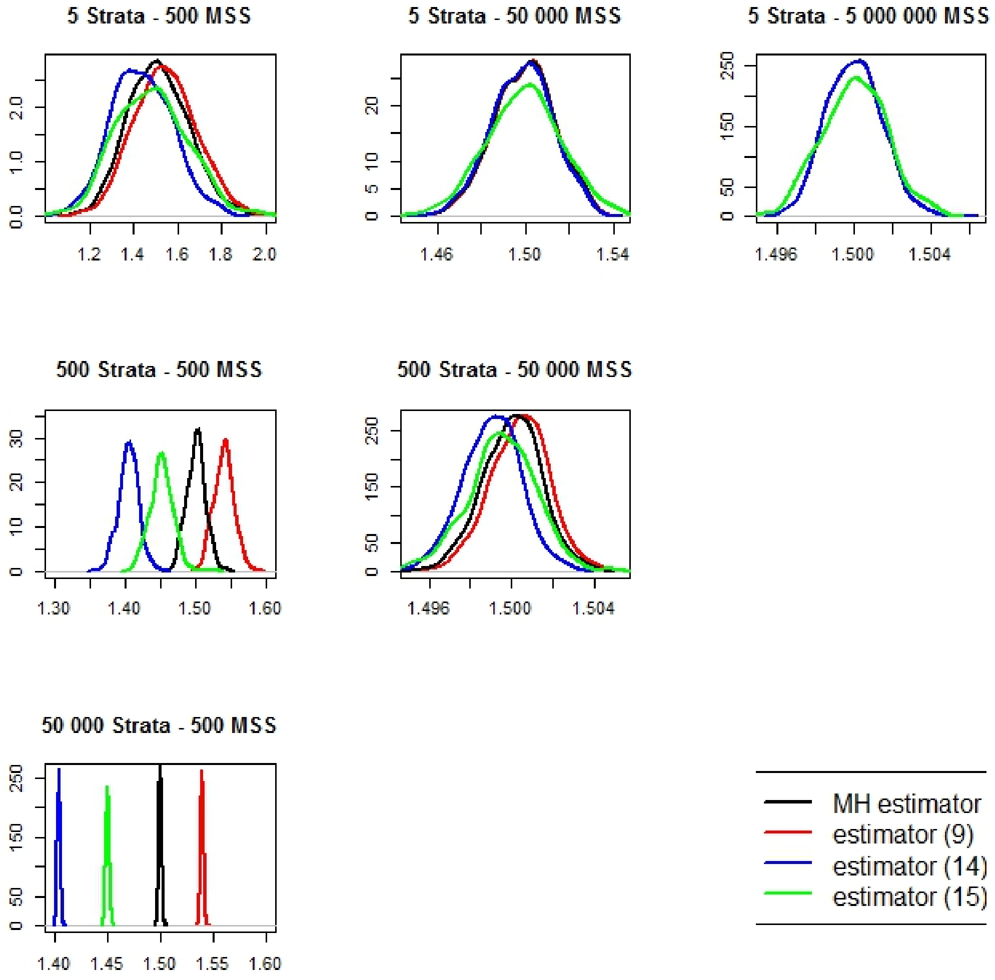

Figure 1: Simulation: Density plots for the Mantel-Haenszel estimator and estimators (9), (14) and (15). (MSS = Mean Stratum Size)

Table 16: Case Study: General 2×2 table

	Diabetes +	**Diabetes -**	Total
Female	a_i	b_i	n_{1i}
Male	c_i	d_i	n_{2i}
Total	m_{1i}	m_{2i}	n_i

Intego procedures were approved by the ethical review board of the Medical School of the University of Leuven (ML 1723) and by the Belgian Privacy Commission (SCSZG/13/079). Many general practices applied for inclusion in the registry. Before approval, the registration performance was checked using algorithms between all participants. Only those with an optimal performance were included. All participating general practices need to routinely record all new diagnoses, drug prescriptions, laboratory results and patient information. They use universal codes; diagnoses are classified using ICPC 2 codes (International Classification of Primary Care) and the WHO's Anatomical Therapeutic Chemical (ATC) classification system for drugs. See also Truyers *et al.* (2014) and the Intego website (http://www.intego.be).

The data set analysed contains information about 338,581 patients, listing their gender, year of birth, the general practices, and diagnosis recorded with the correct ICPC Code. We chose to make several 2×2 tables with the binary variables diabetes and gender (Table 16). General practices and the year of birth can serve as stratification variables. There are 75 different general practices. The years of birth vary from 1898 to 2015. Table 17 presents the estimates for the common odds ratio and variance according to different stratification variables. In the first row general practices (GP) divided the sample in different strata. For the second row, year of birth (YB) was split into two groups, those who where born before 1950 and those born in 1950 and later. In further rows more splits were made.

The differences between MH and (9) is negligible when there is a large number of strata, but that is not the case otherwise. All estimators are equally easily computed, no matter the number of strata. In the last line of the table, the stratum with people born before 1900 has one zero in its 2×2 table. As we stated earlier and now can see here: the MH is well-defined and gives an estimate, however due to a zero weight this specific stratum is omitted in the calculation. On the other hand, the formally optimal estimator does not yield a well-defined estimate.

Table 17: Intego Data: Common odds ratio and variance estimates: (a) ψ_{MH}: Mantel-Haenszel odds ratio estimate, for which also the 95% C.I. is given; (b) $\tilde{\psi}$: odds ratio estimate with (9); (c) $\tilde{\tilde{\psi}}$: odds ratio estimate with (14); (d) $\tilde{\tilde{\psi}}'$: odds ratio estimate with (15); (e) v_R: variance estimate according to Robins *et al.* (1986b) (2); (f) v: variance estimate with (10)

Stratification	# strata	ψ_{MH}	$\tilde{\psi}$	$\tilde{\tilde{\psi}}$	$\tilde{\tilde{\psi}}'$	v_R	v
GP	75	0.967[0.938;0.997]	0.937	0.904	0.944	2.330×10^{-4}	2.524×10^{-4}
YB[1]	2	0.921[0.892;0.951]	0.954	0.916	0.951	2.303×10^{-4}	2.286×10^{-4}
YB[2]	4	0.930[0.901;0.960]	1.108	0.907	0.895	2.294×10^{-4}	2.367×10^{-4}
YB[3]	8	0.928[0.899;0.958]	-	-	-	2.305×10^{-4}	-

[1] split by 1950

[2] split by 1920, 1950, 1980

[3] split by 1900, 1920, 1935, 1950, 1965, 1980, 2000

6 Ramifications and Concluding Remarks

To study associations between binary variables, incorporating classification, the common odds ratio is often used. The odds ratio estimator of Mantel & Haenszel (1959) well established in epidemiology. It is a weighted estimator, combining information across strata, that need not be of the same size. It follows neither from the likelihood principle, nor from optimally weighted considerations. We therefore studied it against the background of an optimally weighted estimator and two variations thereof. We noted that, while there is no complete sufficient statistic, there nevertheless is a uniformly optimal weighted estimator. Unlike a full likelihood estimator, both the MH as well as the optimal estimator, and the variance thereof, are easy to calculate.

The MH, in spite of its somewhat *ad hoc* nature is very efficient for large datasets, and in more cases well-defined when cell counts are small or even zero. It is fair to say the the optimal estimator has a somewhat easier and more intuitive variance expression but, as we have already seen, both are easy to compute.

In some cases when datasets are huge, the optimal estimator is somewhat more efficient. However, we should bring forward two reservations. First, this is not the case when the number of strata becomes equal to or larger than the mean stratum size the estimator fails. Second, with really huge samples, a slight amount of efficiency loss is usually not an issue.

Categorical variables may take on more than 2 levels and also here associations can be of scientific interest. This suggests that one might also consider a possible extension to $I \times J \times K$ tables as is done for the MH (Agresti, 2002, pp. 295).

In conclusion, the MH retains its practical and theoretical attraction, even when compared with formally optimal estimators, and for a large class of settings it will remain the estimator of choice.

For very large datasets, which are increasingly prevalent nowadays, the optimally weighted estimator has some advantages, namely smaller MSE whilst retaining computational efficiency. The study of the weighted estimators helps us understand better the unique position of the MH among statistical estimators.

Acknowledgments

Financial support from the IAP research network #P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. The research leading to these results has also received funding from the European Seventh Framework programme FP7 2007–2013 under grant agreement Nr. 602552. We gratefully acknowledge support from the IWT-SBO ExaScience grant. Intego is funded on a regular basis by the Flemish Government (Ministry of Health and Welfare).

References

Agresti, A. (2002). *Categorical Data Analysis.* Second edition. New Jersey: Wiley series in probability and statistics.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, **15**, 377–380.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research*. Volume 1 - The analysis of case-control studies. Lyon: International Agency for Research on Cancer.

Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.

Flanders D.W. (1985). A new variance estimator for the Mantel-Haenszel odds ratio. *Biometrics*, **41**, 637–642.

Hauck, W.W. (1979). The Large Sample Variance of the Mantel-Haenszel Estimator of a Common Odds Ratio. *Biometrics*, **35**, 817–819.

Department of general practice, KU Leuven. Intego-project. [Online]. 2011 [cited 2018 04 07]; Available from: URL:http://www.intego.be.

Kuritz, S.J., Landis, J.R. and Koch, G.G. (1988). A general overview of Mantel-Haenszel methods: application recent developments. *Annual Reviews Public Health*, **9**, 123–160.

Hermans, L., Molenberghs, M., Aerts, M., Kenward, M.G and Verbeke, G. (2018). A tutorial on the practical use and implication of complete sufficient statistics. *International Statistical Review*, **00**, 000–000.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Clusters with unequal size: maximum likelihood versus weighted estimation in large samples. *Statistica Sinica*, **28(3)**, 1107–1132.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Fast, closed-form, and efficient estimators for hierarchical models with AR(1) covariance and unequal clusters sizes. *Communications in Statistics: Simulation and Computation*, **47(5)**, 1492–1505.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22–4**, 719–748.

Molenberghs, G., Verbeke, G. and Iddi S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics and probability letters*, **81**, 892–901.

Pattanayak, C.W., Rubin, D.B and Zell, E.R. (2012). A potential outcomes, and typically more powerful, alternative to Cochran-Mantel-Haenszel". *SSRN*.

Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **(42(2)**, 311–232.

Robins, J., Greenland, S. and Breslow, N. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology*, **124(5)**, 719–723.

Truyers, C., Goderis., G., Dewitte, H., vanden Akker, M. and Buntinx, F. (2014). The Intego database: background, methods and basic results of a Flemish general practice-based continuous morbidity registration project. *BMC Medical Informatics and Decision Making*. **14(1)**, 48.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of human genetics*, **19**, 251–253.

A Calculations of the Optimal Weights

Obtaining optimal weights is standard, but nevertheless useful to briefly review here for our purposes. We seek to minimize the variance, $\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \alpha_i^2 \text{var}(\psi_i)$. Write $v_i = \text{var}(\psi_i)$ so $\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \alpha_i^2 v_i$, and define the objective function:

$$Q = \sum_{i=1}^{N} \alpha_i^2 v_i - \lambda \left(\sum_{i=1}^{N} \alpha_i - 1 \right), \tag{16}$$

with λ a Lagrange multiplier.

To properly calculate the weights, the first derivative of Q with respect to weights α_i are taken and equated to zero:

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_i} &= 2\alpha_i v_i - \lambda = 0, \\ 2\alpha_i v_i = \lambda \quad &\Rightarrow \quad \alpha_i = \lambda \frac{1}{2v_i}. \end{aligned} \tag{17}$$

By summing both sides, the left hand side is equal to 1 and an expression for λ is obtained:

$$1 = \sum_i \alpha_i = \frac{\lambda}{2} \sum_{i=1}^{N} \frac{1}{v_i} \Rightarrow \lambda = \frac{2}{\sum_{i=1}^{N} \frac{1}{v_i}}. \tag{18}$$

Plugging (18) into (17), the weights are:

$$\alpha_i = \frac{v_i^{-1}}{\sum_j v_j^{-1}}. \tag{19}$$