SIMULATION COMPARISON OF STATISTICAL METHODS USED IN ASSESSING VACCINE EFFICACY IN VETERINARY BIOLOGICS

Kenny Wakeland
Iowa State University, wakeland@iastate.edu

Brian Fergen
Boehringer Ingelheim Vetmedica Inc., brian.fergen@boehringer-ingelheim.com

SIMULATION COMPARISON OF STATISTICAL METHODS USED IN ASSESSING VACCINE EFFICACY IN VETERINARY BIOLOGICS

Kenneth Wakeland, Iowa State University, Department of Statistics

Brian J Fergen, Boehringer Ingelheim Vetmedica, Inc

Abstract:

In veterinary biologics, clinical studies conducted to support the licensure of a vaccine generally include a demonstration of efficacy in the species of interest. Typically, these studies are designed to assess a vaccine's ability to prevent or mitigate clinical disease. Study designs utilize two or more treatment groups, and often incorporate blocking structure restrictions to accommodate animal housing or litter-related effects. When assessing a vaccine's ability to prevent clinical disease, the prevented fraction (PF), a function of the group proportions of affected animals, is often utilized. Typically the sample size per treatment group is limited, and each block is represented by only a few experimental units per treatment group. Thus, it is a common occurrence for group proportion estimates to be 0 or 1 at the block level. Typical methods utilized in analyzing study data include generalized linear mixed model with delta method for confidence interval (GLMM), Cochran-Mantel-Haenszel (CMH) and Gart & Nam (GN). Through simulation, we compare the performance characteristics (power, bias, coverage) of these methods for a range of study designs, sample sizes and PF values, including an assessment of type 1 error rates. Simulation results suggest CMH generally performs well whereas the GN can perform poorly with regards to type 1 error. GLMM performance varies, depending on the simulated situation. Further, upon closer investigation of GN simulated results, it was determined that the method does not always result in a unique solution.

Keywords: Prevented Fraction, binomial response,

1 Introduction

In veterinary biologics, clinical studies conducted to support the licensure of a vaccine generally include a demonstration of efficacy in the species of interest. Typically, these studies are designed to assess a vaccine's ability to prevent or mitigate clinical disease. Prevention generally is assessed through evaluation on an individual animal basis a binary outcome associated with a case definition of disease which dichotomizes outcomes into the categories of affected and unaffected. Binary outcomes are aggregated across groups to estimate the proportion of affected

animals in each group. When conducting the study to achieve a license in the USA, the USDA suggests the Prevented Fraction (Reference VSM 800.202) be estimated. The Prevented Fraction (PF) can be found using an estimate of the relative risk. Specifically, $PF = 1 - \phi$, where $\phi = p_1/p_0$ is the relative risk. Information regarding the precision of the estimate is captured in the form of a confidence interval. A conclusion that a statistically significant vaccine effect has been demonstrated results when the confidence interval does not include 0.

Efficacy studies often utilize two treatment groups, where one group will be a control group used to assess the overall virulence of the challenge and the second group measures the effect of administration of a vaccine. Routinely, blocking structure restrictions to accommodate animal housing or litter-related effects are necessary, and generally need to be considered in the statistical method utilized to estimate the PF and CI. Due to the magnitude of effectiveness offered by most veterinary vaccines and the robustness of challenge virulence in the control group, the sample size per treatment group is typically small, and each block is represented by a limited number of experimental units per treatment group. Thus, it is a common occurrence for group proportion estimates to be 0 or 1 at the block level. Statistical methods utilized necessarily must be able to successfully address these situations. Methods historically utilized in analyzing these study data include generalized linear mixed model with delta method for confidence interval (GLMM), Cochran-Mantel-Haenszel (CMH) and Gart & Nam (GN). Due to the general lack of knowledge on how the performance characteristics for typical vaccine studies, there is obvious value in an evaluation of the methods via simulation, which is reported herein. In this simulation, we assess power, bias, coverage and nominal type 1 error rates for these methods using experimental designs, sample sizes and PF values reflective of typical efficacy studies in the US veterinary biologics industry.

2 Methods

This section describes the methods used to obtain estimates of the PF and the corresponding confidence intervals. The first method to be considered is the method proposed by Gart and Nam (1988).

2.1 Gart and Nam

The Gart and Nam method begins by assuming the data arise from two independent groups that follow binomial distributions,

$$X_j \sim Bin(n_i, p_i);\ \ j = 0,1.$$

In this paper, group 0 will denote the control group and group 1 will denote the vaccine group. The end goal of this method is to produce an estimate and confidence interval for the relative risk ratio,

$$\phi = \frac{p_1}{p_0}. \tag{1}$$

Using (1), p_1 can be reparameterized as $p_1 = \phi p_0$. This leads to the log likelihood

$$L(\phi, p_0) = \log\left[\binom{n_1}{x_1}\binom{n_0}{x_0}\right] + x_1 \text{logit}(\phi p_0) + n_1 \log(1 - \phi p_0) + x_0 \text{logit}(p_0) + n_0 \log(1 - p_0) \tag{2}$$

Using (2), the score functions are

$$S_\phi(\phi, p_0) = \frac{\partial L(\phi, p_0)}{\partial \phi} = \frac{x_1 - n_1 p_1}{q_1 \phi}, \tag{3}$$

$$S_{p_0}(\phi, p_0) = \frac{\partial L(\phi, p_0)}{\partial p_0} = \frac{x_0 - n_0 p_0}{q_0} + \frac{x_1 - n_1 p_1}{q_1}, \tag{4}$$

where $q_0 = 1 - p_0$ and $q_1 = 1 - p_1$. Substituting $p_1 = \phi p_0$ into (4), the maximum likelihood estimate of p_0 , $\tilde{p}_0$, can be obtained from the quadratic equation

$$a\tilde{p}_0 + b\tilde{p}_0 + c = 0, \tag{5}$$

where $a = (n_0 + n_1)\phi, b = -[(x_0 + n_1)\phi + x_1 + n_0]$, and $c = x_0 + x_1$. The MLE of $p_1, \tilde{p}_1$, is obtained for a given value of ϕ as $\tilde{p}_1 = \phi \tilde{p}_0$. Using the Score function defined in (3), Gart and Nam give an estimate for the variance of (3) based on (Bartlett, 1953)

$$var\left[S_\phi(\phi, \tilde{p}_0)\right] = \frac{1}{\phi^2 u(\tilde{p}_0, \tilde{p}_1)}, \tag{6}$$

$$u(p_0, p_1) = \left(\frac{q_0}{n_0 p_0} + \frac{q_1}{n_1 p_1}\right).$$

With (3) and (6), the approximate $1 - \alpha$ confidence limits can be found using

$$\frac{\left(S_\phi(\phi, \tilde{p}_0)\right)^2}{var\left[S_\phi(\phi, \tilde{p}_0)\right]} = \frac{(x_1 - n_1 \tilde{p}_1)^2}{\tilde{q}_1^2 [u(\tilde{p}_0, \tilde{p}_1)]^{-1}} = z_{\alpha/2}^2. \tag{7}$$

Taking the square root of (7) gives a normal deviate based on the Score function

$$z(\phi) = \frac{(x_1 - n_1 \tilde{p}_1)}{\tilde{q}_1 [u(\tilde{p}_0, \tilde{p}_1)]^{1/2}} = \pm z_{\alpha/2}. \tag{8}$$

A skewness correction factor for (3) can be found based on (Bartlett, 1953). The skewness correction factor can lead to closer to nominal coverage for certain experimental designs detailed in (Gart and Nam, 1988). The skewness correction facor for (3) is provided by

$$\gamma_1[\phi] = [u(p_0, p_1)]^{3/2} \left(\frac{q_1(q_1 - p_1)}{(n_1 p_1)^2} - \frac{q_0(q_0 - p_0)}{(n_0 p_0)^2}\right).$$

The skewness corrected confidence limits for a $1-\alpha$ confidence interval are the solutions to

$$z_s(\phi) = z(\phi) - \frac{\tilde{\gamma}_1[\phi]\left(z_{\alpha/2}^2 - 1\right)}{6}, \quad (9)$$

where $\tilde{\gamma}_1$ is γ_1 with $\tilde{p}_0$ and $\tilde{p}_1$ substituted for p_0 and p_1.

Up to this point, the assumptions on the data are that there are two independent groups (control and vaccine) and the responses from these groups are binomial distributions. However, it is very common in actual studies to have blocking or stratification associated with housing of animals and/or the multiparous nature of some species. Gart and Nam provide a method to incorporate this stratified structure. It is fortunate the general procedure for fitting data with stratification is similar to fitting data with none.

The assumptions when stratification is present becomes

$$X_{j,i} \sim \text{Bin}\left(n_{j,i}, p_{j,i}\right); j = 0,1; i = 1, \ldots, I;$$

where, group 0 is the control group, group 1 is the treatment group, and I is the total number of strata present. The score function for ϕ for all strata is

$$S_\phi.(\phi, \tilde{p}_0) = \sum_{i=1}^{I} S_i\left(\phi, \tilde{p}_{0,i}\right) = \sum_{i=1}^{I} \frac{x_{1,i} - n_{1,i}p_{1,i}}{\tilde{q}_1\phi}.$$

That is, the score function of ϕ for the entire data set is simply the sum of the score functions of ϕ for each stratum. The variance of this score function is approximated by

$$var\left[S_\phi.(\phi, \tilde{p}_0)\right] = \sum_{i=1}^{I} \frac{1}{\phi^2 u_i\left(\tilde{p}_{0,i}, \tilde{p}_{1,i}\right)}$$

$$u_i\left(\tilde{p}_{0,i}, \tilde{p}_{1,i}\right) = \frac{1}{\left[\frac{\tilde{q}_{0,i}}{n_{0,i}\tilde{p}_{0,i}} + \frac{\tilde{q}_{1,i}}{n_{1,i}\tilde{p}_{1,i}}\right]}.$$

Similar to the non-stratified case, the approximate, uncorrected, $1-\alpha$ confidence limits are given by

$$\frac{S_\phi.(\phi, \tilde{p}_0)}{\sqrt{var\left[S_\phi.(\phi, \tilde{p}_0)\right]}} = z_I(\phi) = \sum_{i=1}^{I} \frac{(x_{1,i} - n_{(1,i)}\tilde{p}_{1,i})\left[\sum_i u_i\left(\tilde{p}_{0,i}, \tilde{p}_{1,i}\right)\right]^{\frac{1}{2}}}{\tilde{q}_{1,i}} = \pm z_{\alpha/2} \quad (10)$$

The skewness corrected $1-\alpha$ confidence limits are

$$\gamma_{1,i}(\phi) = \frac{\sum_{i=1}^{I}\left[\tilde{q}_{1,i}\left(\tilde{q}_{1,i}-\tilde{p}_{1,i}\right)/\left(n_{1,i}\tilde{p}_{1,i}\right)^2 - \tilde{q}_{0,i}\left(\tilde{q}_{0,i}-\tilde{p}_{0,i}\right)/\left(n_{0,i}\tilde{p}_{0,i}\right)^2\right]\left[u_i\left(\tilde{p}_{0,i},\tilde{p}_{1,i}\right)\right]^{-3}}{\left(\sum_i u_i\left(\tilde{p}_{0,i},\tilde{p}_{1,i}\right)\right)^{-3/2}}$$

$$z_{Is}(\phi) = z_I(\phi) - \frac{\tilde{\gamma}_{1,i}(\phi)\left(z_{\alpha/2}^2-1\right)}{6} = \pm z_{\alpha/2} \qquad (11)$$

Equation (10) is used for more than just the confidence limits however. Due to the added complexity of the stratification, it is necessary to use (10) to obtain a point estimate for the relative risk ratio ϕ. This is done by letting $z_I(\phi) = 0$.

The point estimate and confidence interval for the relative risk, ϕ, can be used to find the point estimate, specifically

$$\widehat{PF} = 1 - \hat{\phi}.$$

The confidence interval is given by

$$\left(1-\hat{\phi}_{1-\alpha/2}, 1-\hat{\phi}_{\alpha/2}\right).$$

Solutions for the $z(\phi)$ functions described in this section require the use of an iterative numeric optimization method. This is, in part, due to the fact estimates of ϕ are needed for $\tilde{p}_1$ and $\tilde{p}_2$, which are in turn needed to estimate ϕ. Gart and Nam suggest using the Secant method to perform this task, as such that is the method implemented in both R (package "PF", version 9.5, 2013-08-29) and the SAS macro used in this paper. For more details on exact implementation and proofs of the relations for this method see (Gart and Nam, 1988).

As a final note, it is important to note that in finding solutions to (8), (9), (10) and (11) there is an implicit assumption the functions have a unique solution, and thus are 1-to-1 (and invertible). If such an assumption were not met, there would be multiple confidence limits, at most one of which would be able to have the correct coverage.

2.2 Cochran-Mantel-Haenzel

The second method discussed here is the Cochran-Mantel-Haenzel (CMH) method for finding common relative risks. This method differs slightly from the Gart and Nam method discussed in the previous section, in that, this method requires the data to be stratified. The assumptions for this method are the data have to be a $2 \times 2 \times I$ contingency table, i.e., the data are stratified 2×2 contingency tables. An example of the type of table required for CMH relative risk estimates is displayed in Table 1.

The CMH method is similar to the Gart and Nam method for stratified data in that a sum of stratum level information is used. Specifically, the estimate of the relative risk is

$$RR_{MH} = \frac{\sum_{h=1}^{I} n_{h12} n_{h2}./n_h}{\sum_{h=1}^{I} n_{h22} n_{h1}./n_h},$$

the variance of which is estimated using the Greenland and Robins (1985) variance estimate of $\log(RR_{MH})$

$$\hat{\sigma}^2 = \widehat{var}[\ln(RR_{MH})] = \frac{\sum_h (n_{h1}.\, n_{h2}.\, n_{h\cdot 2} - n_{h12} n_{h22} n_h)/n_h^2}{(\sum_h n_{h12} n_{h2}./n_h)/(\sum_h n_{h22} n_{h1}./n_h)}.$$

This makes the $1 - \alpha$ confidence interval for the PF

$$(1 - RR_{MH} * \exp[z\hat{\sigma}], 1 - RR_{MH} * \exp[-z\hat{\sigma}])$$

This method is implemented in SAS (using PROC FREQ, SAS ver. 9.4) and R (using package "epiR" ver. 0.9-62). For this paper, the SAS implementation is used.

2.3 GLMM (Delta Method)

The last method of estimating PF and the corresponding confidence intervals is a method based upon the use of the delta method with logistic regression. Similarly to the previous two methods, this method estimates the relative risk, which can then be used to obtain estimates and confidence intervals for the PF.

As mentioned, this method begins with a logistic regression model

$$\text{logit}(p_j) = \mu_j; j = 0,1, \tag{12}$$

or in the case of stratified data

$$\text{logit}(p_{j,i}) = \mu_j + B_i; \qquad j = 0,1;\ i = 1, \dots, I, \tag{13}$$
$$B_i \sim N(0, (\sigma_b)^2).$$

Note here that the strata are being treated as a random block in this method. Using either the models described by (12) or (13), SAS (using PROC GLIMMIX) or R (using "glmer") can produce estimates for μ_0 and μ_1 as well as the covariance of these estimates Σ. These can be used in

$$h(p_0, p_1) = \log(p_1) - \log(p_0) = \log[1 + exp(-\mu_0)] - \log[(1 + \exp(-\mu_1)],$$

the log relative risk ratio. The variance of this is estimated using the delta method to be

$$V = (\nabla h)^T \Sigma (\nabla h).$$

This means the confidence interval of the relative risk is

Strata h		X (affected?)	
		0 (no)	1(yes)
Trt Group	1	n_{h11}	n_{h12}
	0	n_{n21}	n_{h22}

Table 1: Example of level h contingency table required for CMH relative risk estimates.

Litter	Group	#affected	total
74	0	4	4
74	1	1	4
116	0	4	4
116	1	1	4
635	0	2	4
635	1	3	4
796	0	4	4
796	1	3	4
801	0	3	4
801	1	1	4
872	0	4	4
872	1	3	4

Table 2: Table of example data used to illustrate each estimation method

Method	PF Estimate	95% Lower Bound	95% Upper Bound
Score	0.455	0.122	0.655
Skew Corrected	0.455	0.254	0.662
GLMM	0.429	0.121	0.629
CMH	0.429	0.126	0.631

Table 3: Results from each method

$$\left(\exp\left[h(p_0,p_1)+z^*\sqrt{V}\right],\exp\left[h(p_0,p_1)-z^*\sqrt{V}\right]\right), \quad (14)$$

where $z^* = z^*_{\alpha/2}$ is the $\alpha/2$ critical value of the Normal distribution. Notice expression (14) does not have any direct reliance on which model is being considered, stratified or non-stratified. This is due to the fact the difference between these two models are contained entirely within the covariance of the parameter estimates, Σ.

3 Example

Table 2 displays data from a realistic example, which is used to illustrate these methods. The example reflects a design with blocking on litter and randomizing of 4 animals/litter to each of two treatment groups. The results from each of the methods above are presented in Table 3. The "Score" and "Skew Corrected" entries in the table refer to the Gart and Nam method. The "Score" entry refers to the uncorrected confidence interval while the "Skew Corrected" entry refers to the skewness corrected confidence interval. The PF values will always be the same for these two entries. The estimates of the PF are consistent across methods, however the confidence intervals are not consistent across each method. The upper 95% confidence limits display a similar level of consistency to that of the PF estimates, while the lower 95% confidence limits do not. The skewness corrected confidence limit is over twice as large as the estimates of the lower limits for the other methods. This could be explained by a difference in estimation methods, but Table 4 shows this is not the case. Table 4 illustrates a difference in the implementation from R to SAS. The R package "PF" that was initially used to fit this data rounds the root of (5) to 8 decimal places, while the SAS macro rounded the root to 16 decimal places. When the root of (5) was rounded to 16 instead of 8, the results are more in line and consistent with the results from the other methods, and the uncorrected confidence limits from the same method. The first line of thought was the $z(\phi)-$ function for this data was very jagged and this was the reason for the discrepancy. While this is true, the jaggedness of the $z(\phi)-$ function was only part of the story.

Figure 1 is a display of the $z(\phi)-$ function for these data. The story the plots displayed in Figure 1 tell is troubling. The plot on the left is the $z(\phi)-$ function when the root of (5) is rounded to 8 decimal places. The plot on the right is the $z(\phi)-$ function when the root is rounded to 16. Both plots are actually points, with no lines connecting them, with ~10^{-5} distance between points. The plot on the right shows a very jagged function from $\phi = (\sim.65, \sim.78)$, where the value of $z(\phi)$ actually jumps from the lower "line" to the upper "line" for small changes in ϕ. The reason the functions using two different rounding options converge to a different value can be seen using the debugging options in both R and SAS. That is, both functions start at the same place, but by the third step, the function that rounds (5) to 16 decimals has left the lower "line" and, in fact, has left the jagged area completely. This seems to call into

	Method	PF Estimate	95% Lower Bound	95% Upper Bound
Rounding: 10^{-8}	Score	0.455	0.122	0.655
	Skew Corrected	0.455	0.254	0.662
Rounding: 10^{-16}	Skew Corrected	0.455	0.0990	0.662

Table 4: Results for Score and Skew Corrected methods when the rounding was changed from 8 to 16.

	# of Strata *(i)*		
Subjects/block/treatment	***k=4***	***k=6***	***k=8***
2	-	-	✓
3	-	✓	✓
4	✓	✓	✓

Table 5: Strata and subjects/strata combinations used in the simulation study

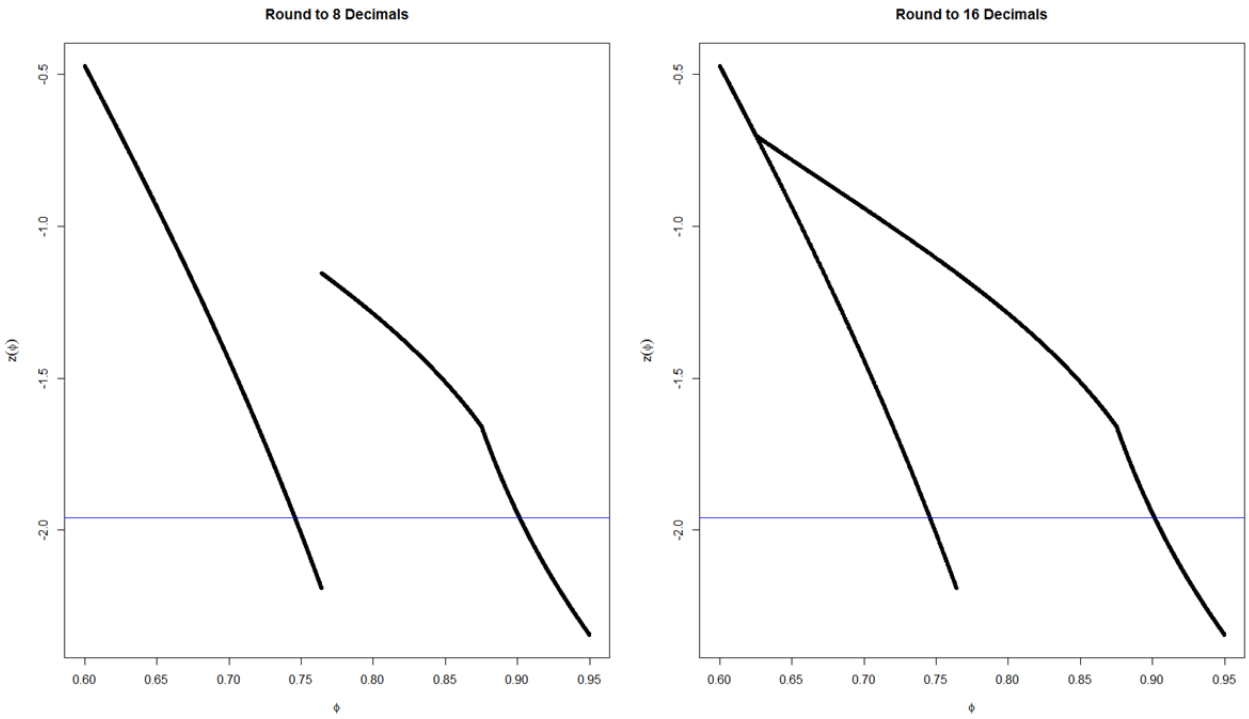

Figure 1: $z_I(\phi)$ − function for the lower confidence interval of ϕ for the data in Table 2

question which rounding value should be used, 16 was chosen initially in SAS (16 digits is the default precision in SAS), the reasoning for rounding (5) to 8 is not clear.

The rounding issue aside though, Figure 1 shows a more troubling aspect of the Gart and Nam method. That is, the $z(\phi)$ −function is not always 1-to-1. This is a serious concern, because as we noted above, if the $z(\phi)$ − function is not 1-to-1, and hence does not have a unique inverse, then it is possible to obtain confidence limits (and point estimates) which are not unique. This leads to the question which limit would produce the correct coverage, and how would the user be able to determine said limit? As a further example of this type of problem with the Gart and Nam method, see Figure 2. This figure displays the $z(\phi)$ −function for a recent vaccine efficacy study. Equation (5) was rounded to 8 decimals, and yet the $z(\phi)$ −function produce looks far worse than the two displayed in Figure 1. There are 4 roots to this function, the algorithm returns the third root ($\phi \sim 1.5$), where the GLMM and CMH method return a result close to the first root ($\phi \sim 0.97$).

4 Simulation Study

The issues surrounding the Gart and Nam method apparent in realistic data examples motivated a more in depth investigation of each method presented so far. The specific investigation presented here is a simulation study in data are simulated from a logit normal distribution of the nature

$$z_{j,i}\ iid\ N(0,1); j = 0,1; i = 1, \ldots, I;$$

$$\text{logit}(p_{j,i}) = \sigma_b * z_{j,i} + \mu_j;$$

where j denotes the treatment group (0 control, 1 vaccine), i denotes the strata or block. Values of the block to block variability, σ_b, used were (0, 0.25, 0.50). The level of the control, μ_0, was held constant at logit(0.90). The levels used for treatment group, μ_1, were logit(0.90), logit(0.60), and logit(0.30).The number of blocks used were 4, 6, and 8, and the subjects per block per treatment used were 2, 3, and 4. However, a minimum number of subjects per study was set at 32, thus some combinations of number of strata/subjects per strata were not used. The combinations used in the simulation study are displayed in Table 5. There are a total of 54 experimental designs proposed, and for each experimental design, 10,000 datasets were simulated.

4.1 Selected Simulation Results

The results for the experimental design that consisted of 8 blocks, 4 subjects/block/treatment, variances 0 and 0.5, and PF=2/3 ($p_{vac} = 0.3$) are displayed in Tables 6 and 7. Table 6 displays the results for the convergence and accuracy of the methods, while Table 7 focuses on the

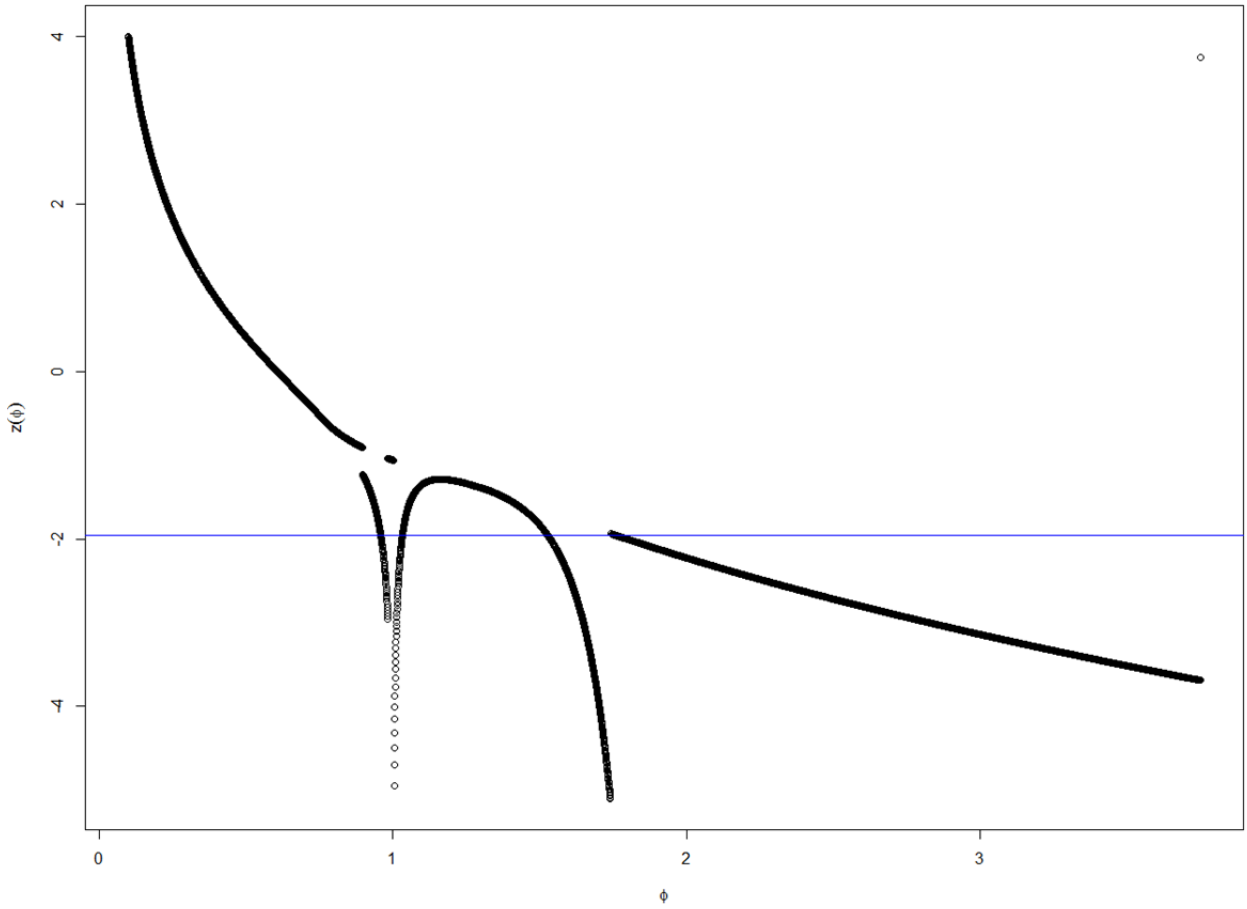

Figure 2: $z_I(\phi)$ for the lower confidence interval of ϕ for study data described in Section 3.

Convergence and Accuracy (Bias)						
	σ_B=0			σ_B=0.5		
Method	**Convergence Rate**	**PF Estimate**		**Convergence Rate**	**PF Estimate**	
		Mean	**Median**		**Mean**	**Median**
CMH	1.000	0.665	0.667	1.000	0.651	0.655
GLMM	0.967	0.668	0.670	0.973	0.656	0.658
Gart and Nam	0.998	0.667	0.670	0.990	0.653	0.657

Table 6: Simulation results for accuracy and convergence for PF=2/3, # of blocks=8, #subjects/block/treatment=4

	σ_B=0		σ_B=0.5	
Method	**Coverage**	**Power**	**Coverage**	**Power**
CMH	0.947	0.9997	0.917	0.9987
GLMM	0.958	0.9995	0.932	0.9978
Score	0.944	0.9940	0.918	0.9821
Skewness Corr	0.946	0.9970	0.921	0.9878

Table 7: Simulation results for coverage and power for PF=2/3, # of blocks=8, #subjects/block/treatment=4

Convergence and Accuracy (Bias)						
	σ_B=0			σ_B=0.5		
Method	**Convergence Rate**	**PF Estimate**		**Convergence Rate**	**PF Estimate**	
		Mean	**Median**		**Mean**	**Median**
CMH	1.000	0.664	0.667	1.000	0.648	0.667
GLMM	0.808	0.664	0.667	0.833	0.650	0.667
Gart and Nam	0.981	0.667	0.678	0.977	0.650	0.658

Table 8: Simulation results for accuracy and convergence for PF=2/3, # of blocks=8, #subjects/block/treatment=2

coverage and power of the methods. It is important to note, for these tables and all tables that follow, the results are conditioned on the method converging. If the method did not converge, the PF estimate and confidence interval was not included in the respective summaries. It is also worth mentioning here the Gart and Nam method is not broken into "Score" and "Skew Corrected" for Table 6, as only the PF estimate and the convergence is being examined. Because the skewness correction only effects the confidence interval, it is possible to simply for this table. Table 7, however, does make the distinction again as this table is displaying characteristics of the confidence interval.

The models perform well and are consistent with each other in terms of accuracy for this scenario. Each method produces PF values that are centered close to the true PF when the block variance is 0. When the block variance increases to 0.5, the distribution of the estimated PF's is biased down slightly. The convergence rate for the models is not as well behaved. The GLMM performs slightly worse than the other two methods, but this can be explained. That is, the GLMM will fail to produce a result when all subjects in one treatment group are affected, or unaffected. In the case where the PF=2/3, the probability of all subjects in one group being affected is approximately 0.034. This corresponds to the convergence rate observed for the GLMM. The few failures to converge of the Gart and Nam method here are related to the secant method failing to find a solution.

The results displayed in Table 7 show the methods are performing well compared to each other, with the CMH and GLMM methods having slightly better power than the Gart and Nam method. The coverage is also close to 0.95 for each method when the block variance is 0, but decrease when the block variance is 0.5. In the situation with $\sigma_B = 0.5$, the GLMM outperforms the other 2 methods.

These results are, for the most part, mirrored when the number of subjects/block/treatment are reduced to 2. These results are displayed in Tables 8 and 9. The only notable deviations are the convergence rate of the GLMM, which still corresponds to the probability of all subjects in a treatment group being affected (~0.185) , and the coverage for all the methods are now very close, with the Gart and Nam method outperforming the other two slightly when $\sigma_B = 0.5$.

The results change dramatically when we consider these same experimental designs, but with a PF=0. These results are displayed in Tables 10 and 11. The convergence rate for the CMH and GLMM methods are now lower than in Table 6, but this can again be explained by the probability of all the subjects of either group being affected. The situation in which all subjects are affected is the only situation in which the CMH fails to produce a result here. The Gart and Nam method appears to be behaving poorly overall when PF=0, however. The convergence is far lower than can be explained by a failure of the secant method to converge, and the bias in the distribution of estimated PF values is much larger than previously observed. These concerns are

	σ_B=0		σ_B=0.5	
Method	**Coverage**	**Power**	**Coverage**	**Power**
CMH	0.955	0.960	0.938	0.938
GLMM	0.954	0.916	0.938	0.870
Score	0.957	0.860	0.948	0.831
Skewness Corr	0.960	0.880	0.950	0.846

Table 9: Simulation results for coverage and power for PF=2/3, # of blocks=8, #subjects/block/treatment=2

Convergence and Accuracy (Bias)						
	σ_B=0			σ_B=0.5		
Method	**Convergence Rate**	**PF Estimate**		**Convergence Rate**	**PF Estimate**	
		Mean	**Median**		**Mean**	**Median**
CMH	0.999	-0.005	0	1.000	-0.004	0
GLMM	0.933	-0.004	0	0.973	-0.004	0
Gart and Nam	0.594	-0.669	0.000	0.593	-0.749	-0.016

Table 10: Simulation results for accuracy and convergence for PF=0, # of blocks=8, #subjects/block/treatment=4

	σ_B=0		σ_B=0.5	
Method	**Coverage**	**Power**	**Coverage**	**Power**
CMH	0.955	0.045	0.950	0.050
GLMM	0.989	0.011	0.982	0.018
Score	0.444	0.556	0.454	0.546
Skewness Corr	0.529	0.471	0.551	0.449

Table 11: Simulation results for coverage and power for PF=0, # of blocks=8, #subjects/block/treatment=4

matched by Table 11 where coverage is not close to the value that is expected. Further investigation of this issue exposes similar issues to the phenomenon observed in the previous section.

Consider the data displayed in Table 12. If the naïve estimate of the PF is used ($\sum_i x_{1,i} / \sum_i x_{0,i}$), the PF would be -0.156, nearly the same as the estimate of the PF for CMH and the GLMM shown in Table 13. However, the result for the Gart and Nam method are not close to that value. The result shown in Table 13 is actually a "failure to converge" error. The macro returned an error code indicating the method failed to converge and a solution could not be found, and the results of the last iteration were returned. Figure 3 provides some insight on the reason for this result. The plot displayed in Figure 3 is the $z_I(\phi)$-function for the data in Table 12, the horizontal line in this figure represents the point estimate of ϕ ($z(\phi) = 0$). It is very clear this function does not have a root at 0, and in fact, is monotone increasing for $\phi > 1.4$. This case is far from unique, and in fact, represents many of the failures to converge when the true PF value was 0. Such situations were still prevalent when the true PF was greater than 0 however. The common thread between these errors seems to be if the observed PF is close to 0, the chances for the Gart and Nam to produce a $z_I(\phi)$-function of this kind increase. It should be noted as well, the plot of $z_I(\phi)$ displayed in Figure 3 is also not invertible, and this, too, is common when these failures occur. Errors such as those described above do help us understand how common situations similar to our example data occur, and unfortunately, they are common when the observed PF gets closer to 0 for these sample sizes.

4.2 Overall Results

Simulation results that aggregated over various variables are displayed in Tables 14, 15 and 16. Table 14 shows a breakdown of the results by PF value, Table 15 displays a breakdown of the results by σ_b value, and Table 16 displays a breakdown of the results by number of strata and subjects per strata. Each PF value and σ_b value, have 180,000 simulated experiments, while each strata/subject breakdown has 90,000 simulated experiments. The results for block-to-block variability seem promising, in that CMH and Gart and Nam seem to perform similarly in the presence of block to block variability as they do when there is no variability between blocks. The GLMM should be best suited to deal with these issues, and in terms of bias and MSE, the GLMM does perform the best, narrowly edging out CMH. The coverage for the GLMM is conservative, on average, however. The convergence rate, coverage and the MSE echo the concerns raised above with the Gart and Nam method.

This concern is mirrored in Table 16. The results for CMH and the GLMM correspond to the expected behavior of these methods when sample sizes are increased, that is, the bias and the MSE have a negative relationship with the sample size. The coverage remains roughly the same for all methods, though as in Table 15.

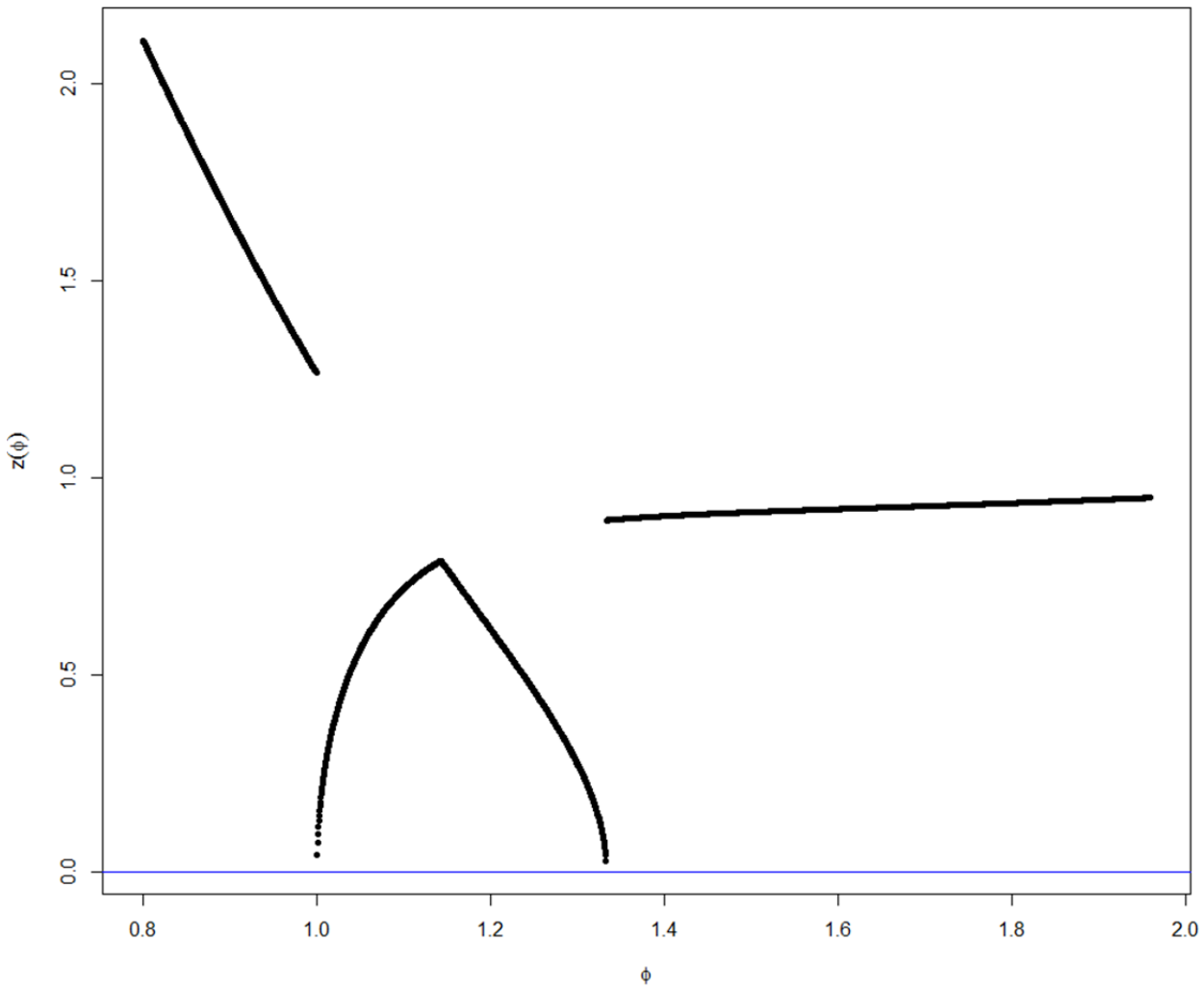

Figure 3: $z_I(\phi)$- function used to estimate the PF for data displayed in Table 9

Strata	Control (x_0)	Vaccine(x_1)	n/group
1	3	4	4
2	3	4	4
3	3	3	4
4	3	3	4

Table 12: Problematic data for the Gart and Nam method

Method	PF	95% Lower Bound	95% Upper Bound
Score	-1.59E14	-1.447	0.168
Skewness Corr	-1.59E14	-227.790	0.162
GLMM	-0.167	-0.636	0.168
CMH	-0.167	-0.636	0.168

Table 13: Results for the data displayed in Table 9

True PF	Method	Convergence Rate	Mean Estimated PF	MSE	Bias	Coverage	Power
0	CMH	0.984	-0.010	0.013	-0.007	0.960	0.040
0	GLMM	0.791	-0.010	0.010	-0.005	0.994	0.006
0	Score	0.499	-0.440	6.888	-0.440	0.443	0.557
0	Skewness Corr	0.501	-0.440	6.857	-0.438	0.531	0.469
1/3	CMH	1.000	0.329	0.019	-0.004	0.940	0.595
1/3	GLMM	0.886	0.321	0.019	-0.012	0.939	0.481
1/3	Score	0.935	0.264	0.816	-0.069	0.943	0.684
1/3	Skewness Corr	0.935	0.263	0.831	-0.070	0.946	0.628
2/3	CMH	0.998	0.658	0.015	-0.009	0.943	0.976
2/3	GLMM	0.884	0.661	0.015	-0.005	0.950	0.943
2/3	Score	0.994	0.660	0.016	-0.006	0.943	0.945
2/3	Skewness Corr	0.994	0.660	0.016	-0.006	0.944	0.951

Table 14: Simulation Results broken down by PF used for simulation

Table 14 better explores the convergence rate, coverage and MSE concerns. The convergence discussed here is the proportion of simulates in which the model produced usable results (that is, no errors were produced). The CMH method did the best in terms of this measure, followed by the GLMM and then the Gart and Nam method. The vast majority of the convergence failures in the GLMM can be explained by the argument above, however, there are some cases when PROC GLIMMIX does not converge but not all subjects in either treatment group were affected. In these cases, PROC GLIMMIX options such as an alternative optimization procedure and/or evaluation over a grid of starting values may increase convergence rates, but was not pursued for this simulation. In addition to the situation described above, the CMH method will fail to produce a result when the denominator (the control group in this case) is entirely unaffected (this does not happen here), or all subjects of the vaccine group are unaffected. These three cases encompass the entirety of the failures to converge of the CMH method.

Refer to Table 14, when the true PF used in the simulation was 2/3, the convergence results, all the results in fact, for each of the methods look similar, and there is not strong evidence to say one method is better than the others. But there is cause for concern when the true PF is 1/3, in that the MSE for the Gart and Nam method is so much higher than the other two methods. The concern is heightened when looking at the results for a true PF of 0. In this case, the MSE is orders of magnitude larger than the other two methods, and the coverage is not close to 0.95. This is paired with the fact the method failed to converge half of the time. These results, combined with the other results in this paper, clearly lay out concerning behavior in the Gart and Nam method.

5 Discussion

When considering all of the results presented in this paper from the simulation study, the CMH method seemed to be the most robust and the method with the fewest drawbacks. This method had the best power when PF=2/3, and was on par with the Gart and Nam, which performed the best, when PF=1/3. However, it is important to note that this was conditioned on the method converging. If the trials where the Gart and Nam method are counted as failures for power, the CMH method would be much higher than the Gart and Nam method. This is assisted by the fact that the CMH method was the most likely to produce usable results. That is, the CMH method has the fewest situations that result in a PF not being estimated, or a confidence interval not being produced.

The GLMM is slightly less attractive than the CMH method in part because the coverage under the null is generally too conservative. The convergence rate for the GLMM was also not as high as the CMH method due to the decreased number of data combinations that result in an estimable

σ_b	Method	Convergence Rate	MSE	Bias	Coverage
0	CMH	0.994	0.014	-0.004	0.954
0	GLMM	0.846	0.014	-0.005	0.965
0	Score	0.811	1.709	-0.109	0.846
0	Skewness Corr	0.812	1.718	-0.109	0.863
0.25	CMH	0.994	0.015	-0.006	0.950
0.25	GLMM	0.852	0.015	-0.007	0.962
0.25	Score	0.809	1.645	-0.115	0.843
0.25	Skewness Corr	0.810	1.645	-0.115	0.861
0.5	CMH	0.995	0.017	-0.009	0.939
0.5	GLMM	0.863	0.017	-0.010	0.952
0.5	Score	0.807	1.855	-0.135	0.833
0.5	Skewness Corr	0.808	1.859	-0.135	0.854

Table 15: Simulation Results broken down by block to block variability used for simulation

# Strata	Subjects per group per Strata	Method	Conver Rate	MSE	Bi	Coverage
4	4	CMH	0.988	0.020	-0.008	0.950
4	4	GLMM	0.769	0.020	-0.012	0.956
4	4	Score	0.781	1.603	-0.090	0.867
4	4	Skewness Corr	0.781	1.603	-0.090	0.891
6	3	CMH	0.992	0.018	-0.007	0.949
6	3	GLMM	0.816	0.018	-0.009	0.960
6	3	Score	0.780	1.678	-0.118	0.851
6	3	Skewness Corr	0.780	1.678	-0.118	0.873
6	4	CMH	0.998	0.013	-0.005	0.945
6	4	GLMM	0.901	0.013	-0.006	0.960
6	4	Score	0.829	1.716	-0.144	0.841
6	4	Skewness Corr	0.829	1.716	-0.144	0.871
8	2	CMH	0.988	0.019	-0.008	0.951
8	2	GLMM	0.777	0.019	-0.011	0.960
8	2	Score	0.764	1.288	-0.050	0.840
8	2	Skewness Corr	0.770	1.320	-0.053	0.836
8	3	CMH	0.998	0.013	-0.006	0.946
8	3	GLMM	0.901	0.013	-0.006	0.963
8	3	Score	0.841	1.783	-0.126	0.823
8	3	Skewness Corr	0.841	1.783	-0.126	0.840
8	4	CMH	1.000	0.010	-0.005	0.944
8	4	GLMM	0.959	0.010	-0.004	0.960
8	4	Score	0.860	2.282	-0.179	0.824
8	4	Skewness Corr	0.860	2.282	-0.179	0.847

Table 16: Simulation results broken down by number of Strata and number of subjects per strata.

PF. Another consideration is that the GLMM model produces subject specific estimates, which requires one to consider carefully the interpretation of the resulting estimate. Here we are using the relative risk estimate for the average stratum, which is not equal to the average relative risk due to the non-linearity of the model and the variability associated with subject-level information and estimator. From Table 15, one can see the bias observed in these simulations is limited, and relatively similar to the CMH method. An additional simulation considering even larger block variance ($\sigma_b^2 = 1, p_{vac} = 0.3$, PF = 2/3, 8 blocks and 4 subjects per treatment per block), resulted in a mean MF estimate of 0.640, which suggests the bias is still limited for relatively large variances. While this is only one small example, it provides additional confidence in the general use of the relative risk of the average stratum. However, we do need to be cognizant of what exactly we are measuring when using this method.

For the simulations parameters considered here (Table 14 specifically), the Gart and Nam method performed relatively well when the PF was 1/3 and 2/3. One cannot, however, ignore the method's inability to approximately hold the nominal type 1 error rate when the PF = 0 (coverage in Table 14). The interpretation of statistical methods, which cannot hold nominal type 1 error rates, is difficult, at best. Here, with 44.3% (Score) and 53.1 % (Skewness Corr) coverage, these methods result in a type error rate ~10 times the nominal of 5%, which would be considered unacceptable in most situations. As mentioned above, the power of this method was actually the highest when PF=1/3, but the bias and the MSE were the worst universally. This is due, in part, to the fact that the $z_I(\phi)$-function can fail for any value of the true PF, it is simply more likely to occur when the true PF value is small. This also leads to the situation where the Gart and Nam had the highest chance to not produce a result. There is also the issue of the fact the $z_I(\phi)$-function is not invertible in general. This can, and will, lead to situations where the method would not be able to produce a unique confidence limit and/or point estimate for the PF.

The CMH and GLMM seem to be the easiest methods to use, as the implementation of each is largely built into SAS/R. The CMH method is an option implemented in PROC FREQ and in the epiR package. The difficult part of the GLMM (the Generalized Linear Mixed Model) is implemented in SAS using PROC GLIMMIX and in R using glmer. The construction of the confidence interval is a straightforward application of the delta method, which can be readily programmed in SAS or R. For the Gart and Nam method, user options include programming the method in SAS, or through the PF package in R.

In conclusion, the Gart and Nam method has many challenges and issues concerning its operating characteristics. The issues described in this paper have lead us to conclude that either the GLMM or CMH methods would be more appropriate to analyze data for which the PF and a confidence interval need to be estimated.

References

1. Bartlett, M. S. (1953). “Approximate confidence intervals, II. More than one unknown parameter.” Biometrika 40, 306-17.
2. Clifford, John R. to Veterinary Services Leadership Team Directors, Centers for Veterinary Biologics Biologics Licensees, Permittees, and Applicants. “General Licensing Considerations: Efficacy Studies for Prophylactic and Therapeutic Biologics.” Veterinary Services Memorandum NO. 800.202, October 24, 2014.
3. Gart, John J., and Nam Jun-mo. "Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness." *Biometrics* 44, no. 2 (1988): 323-38.
4. Greenland, S. and Robins, J. M. (1985), "Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," ***Biometrics***, 42, 311–323.
5. SAS Institute Inc. “PROC FREQ: Cochran-Mantel-Haenszel Statistics” http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_freq_a0000000666.htm (accessed February, 1, 2016).