Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 16

5-1-2006

Properties Of Bound Estimators On Treatment Effect Heterogeneity For Binary Outcomes

Edward J. Mascha
The Cleveland Clinic Foundation, maschae@ccf.org

Jeffrey M. Albert
Case Western Reserve University, jma13@case.edu

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 181-193

1538 – 9472/06/$95.00

Properties Of Bound Estimators On Treatment Effect Heterogeneity For Binary Outcomes

Edward J. Mascha
Department of Quantitative Health Sciences
The Cleveland Clinic Foundation

Jeffrey M. Albert
Department of Epidemiology and Biostatistics
Case Western Reserve University

Variability in individual causal effects, treatment effect heterogeneity (TEH), is important to the interpretation of clinical trial results, regardless of the marginal treatment effect. Unfortunately, it is usually ignored. In the setting of two-arm randomized studies with binary outcomes, there are estimators for bounds on the probability of control success and treatment failure for an individual, or the treatment risk. Here, those bounds were refined and the sampling properties were assessed using simulations of correlated multinomial data via the Dirichlet multinomial. Results indicated low bias and mean squared error. Moderate to high intraclass correlation (ICC) and large numbers of clusters allow narrower confidence interval widths for the treatment risk.

Key words: Blocked or clustered data, bounds, causal effects, Dirichlet multinomial, intraclass correlation, marginal treatment effect, randomized trial, potential outcomes, treatment effect heterogeneity, unit-treatment interaction.

Introduction

In randomized clinical trials comparing an experimental treatment (T) to a control (C), the focus is usually on the marginal treatment effect, (i.e., mean causal effect) estimated by the difference in means or the difference in the proportion having a successful outcome. Unfortunately, the amount of variability of the individual causal effects is usually ignored. Recent work has seen the development of bounds on a treatment effect heterogeneity parameter for binary outcomes (Gadbury, Iyer, & Albert, 2004; Albert, Gadbury, & Mascha, 2005). The latter provided bound estimates and confidence intervals in the case of blocked binary outcomes. However, no study has been yet conducted to evaluate the properties and practicality of these methods.

Edward J. Mascha, Ph. D., is an Assistant Staff Biostatistician. His interests include causal effects and correlated data methods. Email him at maschae@ccf.org. Jeffrey M. Albert, is an Assistant Professor of Biostatistics. His research interests include causal inference. Email him at jma13@case.edu.

Treatment effect heterogeneity (TEH), also called unit-treatment interaction (Gadbury & Iyer, 2000) or subject-treatment interaction (Gadbury, Iyer, & Allison, 2001), is the amount of variability in the causal effect of T versus C on some outcome Y. The causal effect for an individual is defined as the difference in the individual's potential outcomes (Neyman, 1923; Rubin, 1974; 2000) on T and C, respectively. This is an unobservable latent variable since only one of the two potential outcomes may be observed for an individual. For example, consider a binary outcome scenario with success proportions of 0.50 and 0.30 for treatments T and C, respectively, giving a marginal treatment effect of 0.20. With these marginals, the minimum possible TEH would be that no patients who succeed on C would fail on T, implying that 0.20 of the patients would fail on C and succeed on T. With the same marginals, the maximum possible TEH would be that 0.30 of patients would succeed on C but fail on T, and that 0.50 would fail on C but succeed on T.

Thus, in the case of a binary outcome for two treatments, individuals fall into a category based on their potential outcomes: (1) failure on both T and C, (2) success on T and C, or (3) success on one but not the other. The

probabilities of membership into each of these categories are denoted as π_{00}, $\pi_{01}, \pi_{10}, \pi_{11,}$ where indices indicate response (1=success, 0=failure) to T and C, respectively. The probability of doing worse on a new treatment (T) than on standard treatment (C), $\pi_{01,}$ may be understood as the treatment risk because patients would not expect to do worse on the new treatment. Although this quantity is typically overlooked in analyses of clinical trials, it would be of potential interest for both individual treatment decisions and the understanding of the population impact of treatment.

Albert, Gadbury and Mascha (2005, AGM) provided bounds and bound estimators for the treatment risk π_{01} (referred to by AGM as π_2). However, the AGM bounds cannot be reliably used in practice until their sampling properties have been assessed. Such is the purpose of this article.

Background

Gadbury and Iyer (2000) derived bounds for the probability of an unfavorable individual treatment effect where the outcome is continuous; for example, an individual doing better (higher value) on control than on treatment. They assumed a trivariate normal distribution between the potential outcomes on treatment X and control Y, and a covariate Z which is measured on all patients. Such methods are not easily applicable to binary outcomes because of the difficulty in specifying a meaningful multivariate distribution for the binary setting.

New methods are available, however, to estimate bounds on treatment effect heterogeneity for binary outcomes. These include simple bounds and bounds which make use of clustering. Based on the fact that $\pi_{11,}\pi_{00,}\ \pi_{10,}\ \pi_{01}$ sum to 1.0 and that $\pi_{10} - \pi_{01} = \pi_T - \pi_C$; Gadbury, Iyer, and Albert (2004), which is referred to as GIA, derived simple bounds for π_{01} such that

$$\max(0, \pi_C - \pi_T) \equiv L_S \leq \pi_{01} \leq U_S \equiv \min(1-\pi_T, \pi_C) \quad (1)$$

For example, with true marginal successes π_T = .80 and π_C =.70, simple bounds for π_{01} are (0, .20), and by substituting π_T for π_C and visa versa, the simple bounds for π_{10} are (.10, .30). The marginal proportions π_T and π_C have a large effect on the possible range of unit-treatment interaction in the binary outcome case. A proportion close to 0 or 1 greatly limits the range of TEH, and so allows tighter bounds on the parameters of interest. When neither of the marginals is close to 0 or 1, there is a wider range of possible heterogeneity, and therefore greater opportunity for narrowing through more refined methods.

GIA also give more refined bounds on π_{01}, first using a matched-pairs design in which one member of a pair is randomly assigned to receive treatment and the other member receives control. They construct bounds which narrow as the quality of the matching improves. Further, they consider an extended matched-pairs design, in which some pairs are randomized to either both treatment or both control, which allows the refined bounds to be estimated.

Gadbury, Iyer, and Albert (2004) defined the probability that a treatment unit fails ($Y_T(u_1)=0$) and the matched control unit has success ($Y_C(u_2)=1$), i.e., control beats treatment, or,

$$g_2 = P(Y_T(u_1) = 0, Y_C(u_2) = 1)$$

where u_1 and u_2 are two members of a matched pair. GIA also define h_T and h_C as probabilities of success for both members of a pair of randomly chosen matched treated or control units, respectively, such that

$$h_T = P(Y_T(u_1) = 1, Y_T(u_2) = 1)$$

and

$$h_C = P(Y_C(u_1) = 1, Y_C(u_2) = 1)$$

Higher h_T and h_C indicate better matching and will lead to tighter bounds. Lower and upper bounds for π_{01}, with the "*B*" subscript referring to the blocked (in the present case, the extended matched pairs) design, are as follows:

$$L_C \equiv \text{Max } (0, g_2 - \min(\pi_T - h_T, \pi_C - h_C))$$
$$U_C \equiv \text{Min } (1, g_2 + \min(\pi_T - h_T, \pi_C - h_C)) \quad (2)$$

The bounds for π_{01} (equation 2) were derived by first expressing g_2, h_T and h_C as functions of the underlying parameters of interest, and then adding terms to the expression for g_2 so that the resulting form consisted of quantities for which one has estimators.

In the latest development, Albert, Gadbury, and Mascha (2005, AGM) used bounds with the same form as (2) for π_{01}, but extend definitions to the more general blocked or clustered design. That is, the pair of individuals u_1 and u_2 in the definitions of g_2, h_T and h_C, is now considered as belonging to the same cluster. In many cases this is more realistic than the matched or extended pairs design. Blocks can be created post-hoc. Good blocking or matching gives narrower bounds.

AGM provide non-parametric estimators of the bounds in (2). Each represents a proportion with the given outcome combination, and is estimated as the ratio of the sum across clusters of the number of pairs observed with the given outcome combination to the number of pairs with the given treatment assignments. For example,

$$\hat{g}_2 = \frac{\sum_j n_{C1j} n_{T0j}}{\sum_j n_{Cj} n_{Tj}},$$

is the estimator for g_2, and is the proportion of observed pairs with treatment failure and control success out of the total number of possible treatment-control pairs. Substitution into (2) yields estimated cluster bounds $\hat{L}_B$ and $\hat{U}_B$. AGM (equations 6 through 11) give variances and covariances for estimators of the lower and upper bounds on π_{01} and for their components. Refer to their article for details on the formulae, which are quite extensive.

In this study, the AGM estimators for bounds on π_{01} are first refined. Then, through simulations their statistical properties, including bias, variance, MSE, and coverage are evaluated. Because the AGM bound estimators depend on clustering in the data, a simulation method that allows specification of the intraclass correlation (ICC) as well as the underlying probabilities has been devised. Simultaneous confidence intervals for the lower and upper bounds are shown to provide at least 1-α coverage of π_{01}, the real parameter of interest. Properties are shown to depend on degree of ICC, TEH, marginal success, number of clusters, and sample size.

Methodology

First, a refinement to the AGM bounds is proposed, and then the Dirichlet-multinomial (DMN) is introduced as the model for the potential outcomes. Finally, the treatment effect heterogeneity scenarios and simulation methods used to assess statistical properties of the estimators for bounds on π_{01} and their components are outlined.

Refinement to AGM Bounds

With good blocking, the AGM cluster bounds in (2) are narrower than the simple bounds (1) on π_{01}. However, it can be shown that the cluster bounds are the same or wider than the simple bounds when subjects are independent from each other (and thus, $h_T = \pi_T^2$ and $g_2 = (1-\pi_T)\,\pi_C$), which would occur if the matching or clustering were at random or non-existent. Therefore, a modification of the AGM cluster bounds to be the narrower of the simple and AGM cluster bounds is proposed, such that:

$$L_{MC} \equiv Max(L_S, g_2 - \min(\pi_T - h_T, \pi_C - h_C))$$
$$U_{MC} \equiv Min(U_S, g_2 + \min(\pi_T - h_T, \pi_C - h_C)) \quad (3)$$

With random matching, the modified AGM cluster bounds (MAGM) and simple bounds are identical, and the cluster bound width will always be at least as narrow as the simple bound width, sometimes significantly narrower, depending on the TEH scenario, the marginals, and the amount of clustering.

Property assessment

In order to assess the statistical properties of the bound estimators for $\pi_{01,}$ a model of the underlying (i.e., latent) correlated multinomial data was needed, where each unit or subject belongs to one of the four potential outcome categories (C_{00}, C_{01}, C_{10}, C_{11}), indexed by the latent response to treatment and control, respectively, with probabilities $\pi_{00,}$ $\pi_{01,}$ $\pi_{10,}$ π_{11}, and where units are correlated within clusters. Various approaches to modeling correlated multinomial data have been used (Gange, 1995, Morel & Nagaraj, 1993, Banergee & Paul, 1999). Mosimann (1962) and Brier (1980) extol the Dirichlet multinomial (DMN) distribution, also called the multivariate beta-binomial distribution, as a natural way to model over-dispersed multinomial data. The DMN is used because it also allows direct specification of the intra-class correlation and there is no need to assume an underlying continuous distribution of the data. It is less computationally intensive than some of the other methods and can therefore be used with large numbers of clusters and units per cluster, r, where the method of Gange (1995), for example, cannot.

It is assumed that each unit latently falls into one of the four population categories with the corresponding probabilities $\pi_{00,}$ $\pi_{01,}$ $\pi_{10,}$ π_{11}, denoted as the vector π. Each cluster's set of probabilities deviates randomly from the underlying vector according to the Dirichlet distribution and the counts within each cluster are independent multinomial data conditional on the realized cluster probabilities. The unconditional counts in the 4 categories are distributed as DMN, or $DMN_4(n,\pi,k)$, where *k* is a structural parameter related to the ICC, the correlation among units within the same cluster and category, such that *k*= (1-ICC)/ICC, and so ICC=1/ (1+*k*). This relationship between *k* and the ICC is used to induce varying levels of correlation among subjects within clusters in the simulations.

The statistical properties of the MAGM and AGM estimators for bounds on π_{01} and estimators for their components (g_2, $H_{T,}H_{C,}$ π_T and π_C) were evaluated under five treatment effect heterogeneity (TEH) scenarios (Table 1). Scenarios are distinguished by the level of TEH (low, medium or high value of π_{01} for the given marginals) and the marginal success proportions π_T and π_C: one marginal close to zero (π_T =.20, π_C =.10) or both close to .50 (π_T =.45, π_C =.55). Each scenario is also described by the amount of correlation among the potential outcomes on T and C, or ρ_{PO}. This correlation is a function of π_{01} and the marginal success proportions, so that zero ρ_{PO} indicates independence of the potential outcomes, in which case π_{01} and π_{10} are the product of the corresponding marginals, and which may be the most natural case. Negative ρ_{PO} indicates high TEH (π_{01} and π_{10} are higher than under independence) and positive ρ_{PO} indicates low TEH (π_{01} and π_{10} are lower than expected under independence). Within each scenario, the ICC (.15, .50, and .85), the total sample size *N* (600, 3000), and the number of clusters *C* (20, 40, and 100) are varied to assess the effect of each factor on the estimator properties.

A set of simulations was conducted for each TEH scenario from Table 1, for each variation of ICC, total sample size, and number of clusters. For each cluster *i* , Dirichlet random deviates $p_1^{(i)},\ldots,p_4^{(i)}$ were formed of success probabilities from the underlying vector π as the ratio of random gamma deviates over the sum of the associated four gamma deviates (Jensen, 1998), where subscripts 1, …, 4 indicate the four population categories C_{00}, C_{01}, C_{10}, C_{11}, respectively. The parameter for each of the four gamma deviates is the clustering parameter *k* times the probability of the associated underlying population category. Next, *n* units (where *n*=*N* /C) were randomly sampled from the four population categories according to a multinomial distribution with probabilities $p_1^{(i)},\ldots,p_4^{(i)}$ for the i^{th} cluster. Each unit within each cluster was randomly assigned to have either the response to Y_T or Y_C observed. Finally, the estimated bounds (and estimated bound components) for π_{01}, plus individual and simultaneous (lower, upper bound) confidence intervals for the bounds were calculated. This was repeated 1,000 times for each scenario combination (each particular scenario, sample

size, ICC and number of clusters combination) and summarized across simulations.

For the AGM and MAGM bound estimators and their components within each scenario, the expected value (mean over 1,000 simulations), bias, true variance (variance of the estimated values over the simulations), mean estimated variance and mean squared error (MSE) were assessed. Formula-based 95% confidence intervals (CI) and their widths for lower and upper bounds were then obtained. Approximate confidence intervals were calculated using a normal approximation for the distribution of the bound estimators. For example, a 100(1-α) % confidence interval (CI) for the AGM upper bound, U_B, is $\hat{U}_B \pm z_{1-\alpha/2}(\hat{V}(\hat{U}_B))^{1/2}$, where $z_{1-\alpha/2}$ is the (1-α/2) percentile of the standard normal distribution. A CI for the lower bound, L_B, was obtained similarly. Finally, coverage of the true bounds for both the lower and upper bound estimators was obtained.

Simultaneous (i.e., joint) asymptotic (1-α)% confidence intervals intended to have at least a 1- α probability of containing the true population values of both the lower and upper bounds were also obtained. These were formed by the estimated lower 95% CL of the lower bound and the estimated upper 95% CL of the upper bound from the AGM formulae. Because the formed intervals are designed to have the given nominal probability of containing the true bounds on π_{01}, by definition they should have at least as great a probability of containing the true π_{01}, the parameter of interest. Using these intervals, the mean estimated width, the simultaneous estimated coverage of the true bounds, and the estimated coverage of the true parameter $\tilde{\pi}_{01}$ are reported.

For comparison purposes, and because the joint distribution of the lower and upper bounds is not readily available (assumed to be independent in forming the confidence intervals above), joint confidence intervals were also estimated using a bootstrap method which naturally accounts for dependency between the bounds and also allows non-symmetric intervals around the estimators. Bickel and Friedman (1981) proved that the bootstrap can be used to construct confidence intervals for two unknown parameters simultaneously. Horowitz and Manski (2000) use the bootstrap to put bounds on the treatment effect for missing-value data, where either baseline covariates and/or outcomes are missing for some subjects. The same method was used to provide a joint confidence interval for a pair of lower and upper cluster bounds on the parameter $\pi_{01.}$ The goal was to create an interval of the form $[\hat{L} - d_\alpha, \hat{U} + d_\alpha]$, where $\hat{L}$ and $\hat{U}$. An appropriate value of a constant d_α was chosen such that the interval contains the true parameters L and U with probability 1- α asymptotically. The delta was applied non-symmetrically in hopes of achieving even better coverage with equivalent or smaller confidence interval widths as with the formula method.

Results

Tables 2 and 3 report bias, variance and MSE of the MAGM lower and upper bound estimators for two representative scenarios: scenario 1, the combination of low treatment heterogeneity (π_{01}= .01) and marginals close to zero and scenario 5, the combination of high treatment heterogeneity (π_{01}= .40) and marginals close to .50. Bias of the lower and upper bound estimators and their components is consistently low, typically much less than 5% of the expected value of the estimator for low, medium, or high ICC for each scenario assessed. Bias decreases with increasing ICC. Higher ICC increases the mean estimated variance of the lower and upper bound estimators and components and therefore the MSE, given the consistently low bias. As expected, the mean estimated variances and covariances of the bound estimators across simulations using the AGM formulas are also very close to the true variances and covariances for each estimator. Having a larger number of clusters for a fixed ICC and sample size steadily decreases the variance of all estimators and their associated MSEs. Similar properties and relationships were observed for scenarios 2, 3, and 4 (results not shown).

Confidence interval width and coverage results of both the individual and the simultaneous lower and upper bound estimators on π_{01} are given in Tables 4 and 5 for scenarios

1 and 5, respectively, and in Figures 1 (all scenarios, 20 clusters) and 2 (scenarios 1, 3 and 5 for 20, 40 and 100 clusters). As expected from results on the variance of the bound estimators, CI widths for the individual lower and upper bounds were in general much narrower for scenario 1 (Table 4) and scenario 2 (data not shown), where at least one of the marginal success proportions is close to 0 or 1. Mean CI widths for the lower and upper bounds increase substantially as the ICC increases from 0.15 to 0.85, and this is a function of the variance increasing with ICC. Widths decrease substantially with increasing number of clusters (but C=100 also has a larger total N). The MAGM and AGM methods produce very similar or identical simultaneous (lower, upper) bound widths in cases where the ICC is at least 0.50 (Tables 4, 5) or where neither marginal is close to 0 or 1 (Table 5). The MAGM method has widths that are a 0-20% narrower than the AGM for low ICC and marginals close to 0 or 1 (Table 4, ICC=.15).

Joint CI width of the lower and upper bounds is much narrower when either marginal is close to zero, especially with low to moderate ICC (Figures 1 and 2). The average width of the simultaneous intervals narrows by as much as 50% as the ICC increases from 0.15 to 0.85, and this is more pronounced with larger total sample size. The average joint CI width also decreases substantially as the number of clusters is increased within a fixed sample size, particularly when the ICC is 0.50 or 0.85 (Figure 2). Across all of the scenarios assessed, the average width of the joint intervals is only 3-15 percentage points wider than the width of the true bounds. Higher values of π_{01} (and thus higher TEH) for fixed marginals increase the joint CI width (Figure 1).

Coverage of the individual true bounds was between 90% and 100% for both the AGM and MAGM methods in most situations (Tables 4 and 5, columns H and M). Coverage was above 90% under all scenarios when the ICC was 0.15 or when it was 0.50 and with 30 or more clusters (data shown for 40 and 100 clusters). However, it dropped below 90% with the combined scenario of smaller number of clusters (20), marginals closer to zero, and moderate to high ICC. In a few situations with only 10 clusters (not shown), the coverage was as low as 65-70%. With the unlikely ICC of 0.85 and marginals close to zero or one, forty or more clusters were sometimes needed to obtain coverage of at least 90%.

Simultaneous coverage of the true bounds (column O in Tables 4 and 5) is at least 90% in most cases, and often above 95%. It follows a pattern similar to coverage of the individual bounds, being best when the ICC is moderate or low and with a non-trivial number of clusters (20 or more). In most situations, the coverage was close to or slightly better than the worst of the individual lower and upper bound coverages for that scenario. The width of the simultaneous interval was sometimes narrower for the bootstrap method, but the slightly narrower width was usually accompanied by lower coverage of the true bounds. In general, coverage of the true MAGM bounds was better with the variance formula method than for the bootstrap method (as much as 0.15 better) for similar CI width.

Finally, coverage of the unobservable quantity π_{01} using the simultaneous confidence intervals (column P in Tables 4 and 5) is often 100% and nearly always above 95%. It is affected by the ICC, number of clusters, TEH scenario and total sample size with the same pattern as for the simultaneous bounds coverage.

Conclusion

AGM and refined AGM estimators have good statistical properties (low bias, MSE) and can thus be used in practice to estimate bounds for treatment effect heterogeneity with a binary outcome. Moderately or highly clustered data result in narrower confidence intervals for the measure of treatment heterogeneity π_{01}, the probability of treatment failure and control success, which is termed the treatment risk. Higher ICC is preferable because the bounds themselves move considerably closer to the parameter they are bounding, π_{01}, for larger ICC, and this phenomenon leads to narrower confidence interval widths for the simultaneous bounds as well as for $\pi_{01.}$. A moderate or large number of clusters (at least 20) and larger sample size allow more narrow confidence

Table 1. Simulation scenarios used to assess π_{01} bound estimators and components.

Scenario	Marginal Success		Heterogeneity Descriptions		Prob (Y_T=i, Y_C=j)			
	π_T	π_C	TEH	ρ_{PO}[1]	π_{00}	π_{01}	π_{10}	π_{11}
1	0.20	0.10	Low	.58	.79	.01	.11	.09
2	"	"	Med	.00	.72	.08	.18	.02
3	0.55	0.45	Low	.78	.44	.01	.11	.44
4	"	"	Med	.00	.25	.20	.30	.25
5	"	"	High	-.80	.05	.40	.50	.05

Note: [1] = correlation among potential outcomes on T, C

Table 2. Bias, variance and MSE for Scenario #1 (low TEH + marginals near 0).

			PROPERTY					
θ	ICC	# Clusters	$E(\theta)$	$E(\hat{\theta})$	$E(\theta-\hat{\theta})$	$E(\hat{V}(\hat{\theta}))$	$V(\hat{\theta})$	MSE
LB	0.15	20	0.0000	0.0012	0.0012	0.0001	0.0000	0.0000
		40	.	0.0009	0.0009	0.0001	0.0000	0.0000
		100	.	0.0001	0.0001	0.0000	0.0000	0.0000
	0.5	20	0.0000	0.0063	0.0063	0.0003	0.0001	0.0002
		40	.	0.0061	0.0061	0.0002	0.0001	0.0001
		100	.	0.0030	0.0030	0.0001	0.0000	0.0000
	0.85	20	0.0070	0.0149	0.0079	0.0006	0.0005	0.0005
		40	.	0.0162	0.0092	0.0005	0.0003	0.0004
		100	.	0.0107	0.0037	0.0001	0.0001	0.0001
UB	0.15	20	0.1000	0.0998	-.0002	0.0014	0.0010	0.0010
		40	.	0.1008	0.0008	0.0010	0.0006	0.0006
		100	.	0.1004	0.0004	0.0003	0.0002	0.0002
	0.5	20	0.0900	0.0804	-.0096	0.0015	0.0016	0.0017
		40	.	0.0831	-.0069	0.0009	0.0009	0.0009
		100	.	0.0874	-.0026	0.0003	0.0003	0.0003
	0.85	20	0.0340	0.0268	-.0072	0.0010	0.0009	0.0009
		40	.	0.0304	-.0036	0.0007	0.0006	0.0006
		100	.	0.0344	0.0004	0.0002	0.0002	0.0002

Notes:Marginals: π_T = .20 , π_C= .10; P(Y_T=i,Y_C=j): π_{00}= .79 , π_{01}= .01, π_{10}= .11 , π_{11}= .09; Total N=600 (for C=20, 40), N=300 (for C=100); 1,000 simulations per scenario.

Table 3. Bias, variance and MSE for Scenario #5 (high TEH + marginals near 0.5).

			PROPERTY					
θ	ICC	# Clusters	E(θ)	E($\hat{\theta}$)	E($\theta - \hat{\theta}$)	E($\hat{V}(\hat{\theta})$)	$V\left(\hat{\theta}\right)$	MSE
LB	0.15	20	0.0218	0.0377	0.0159	0.0021	0.0014	0.0017
		40	.	0.0334	0.0116	0.0014	0.0010	0.0011
		100	.	0.0266	0.0049	0.0004	0.0004	0.0004
	0.5	20	0.1775	0.1891	0.0116	0.0068	0.0066	0.0068
		40	.	0.1863	0.0088	0.0039	0.0039	0.0040
		100	.	0.1815	0.0040	0.0014	0.0015	0.0015
	0.85	20	0.3333	0.3447	0.0114	0.0106	0.0112	0.0114
		40	.	0.3434	0.0102	0.0058	0.0060	0.0061
		100	.	0.3382	0.0050	0.0022	0.0021	0.0022
UB	0.15	20	0.4425	0.4284	-.0141	0.0022	0.0021	0.0023
		40	.	0.4248	-.0177	0.0015	0.0013	0.0016
		100	.	0.4365	-.0060	0.0005	0.0004	0.0005
	0.5	20	0.4250	0.4084	-.0166	0.0063	0.0062	0.0065
		40	.	0.4082	-.0168	0.0035	0.0035	0.0038
		100	.	0.4192	-.0058	0.0013	0.0013	0.0014
	0.85	20	0.4075	0.3954	-.0121	0.0103	0.0105	0.0106
		40	.	0.3930	-.0145	0.0056	0.0054	0.0056
		100	.	0.4047	-.0028	0.0021	0.0019	0.0019

Notes: Marginals: $\pi_T = .55$, $\pi_C = .45$; $P(Y_T=i, Y_C=j)$: $\pi_{00} = .05$, $\pi_{01} = .40$, $\pi_{10} = .50$, $\pi_{11} = .05$
Total N=600 (for C=20, 40), N=300 (for C=100); 1000 simulations per scenario.

Table 4. CI width and coverage of bounds on π_{01} for scenario 1: Low heterogeneity and marginals near zero.

			Lower Bound(LB)					Upper Bound(UB)					Simultaneous Lower, Upper		
ICC	#C/#U	Meth	True	L95	U95	W	Cov	True	L95	U95	W	Cov	W	Cov	Cπ_{01}
.15	20/30	AGM	.000	.00	.02	.02	1.0	.15	.07	.22	.15	.90	.22	.92	1.0
		MAGM	.000	.00	.02	.02	1.0	.10	.03	.17	.15	.97	.17	.97	1.0
	40/15	AGM	.000	.00	.02	.02	1.0	.15	.09	.21	.12	.93	.21	.95	1.0
		MAGM	.000	.00	.02	.02	1.0	.10	.04	.16	.12	.97	.16	.98	1.0
	100/30	AGM	.000	.00	.01	.01	1.0	.15	.11	.18	.07	.94	.18	.96	1.0
		MAGM	.000	.00	.01	.01	1.0	.10	.07	.13	.07	.98	.13	.99	1.0
.50	20/30	AGM	.000	.00	.03	.03	1.0	.09	.02	.16	.14	.88	.16	.89	1.0
		MAGM	.000	.00	.03	.03	.99	.09	.01	.15	.14	.86	.15	.87	1.0
	40/15	AGM	.000	.00	.03	.03	1.0	.09	.03	.15	.12	.91	.15	.91	1.0
		MAGM	.000	.00	.03	.03	1.0	.09	.03	.14	.11	.89	.14	.89	1.0
	100/30	AGM	.000	.00	.02	.02	1.0	.09	.06	.12	.07	.94	.12	.95	1.0
		MAGM	.000	.00	.02	.02	1.0	.09	.05	.12	.07	.93	.12	.94	1.0
.85	20/30	AGM	.007	.00	.05	.05	.87	.03	.00	.08	.08	.76	.08	.76	.92
		MAGM	.007	.00	.05	.05	.87	.03	.00	.08	.08	.75	.08	.75	.92
	40/15	AGM	.007	.00	.05	.05	.96	.03	.00	.08	.08	.89	.08	.89	.98
		MAGM	.007	.00	.05	.05	.96	.03	.00	.08	.08	.89	.08	.89	.98
	100/30	AGM	.007	.00	.03	.03	.96	.03	.01	.06	.06	.92	.06	.93	1.0
		MAGM	.007	.00	.03	.03	.96	.03	.01	.06	.06	.92	.06	.93	1.0

Legend: Table values are means over 1000 simulations, except for columns labeled 'True' values ICC= Dirichlet multinomial correlation; #C= number of clusters, #U= number of units per cluster AGM=Equation 2.2; MAGM=Equation 2.6; W=width of 95% CI= U95-L95; Cov=coverage; Simultaneous: coverage of both Lb and UB using L95 of LB, U95 of UB; Cπ_{01}: coverage of π_{01} using L95 of LB, U95 of UB

Table 5. CI width and coverage of bounds on π_{01} for scenario 5: High heterogeneity and marginals near 0.50.

			Lower Bound(LB)					Upper Bound(UB)					Simultaneous Lower, Upper		
ICC	#C/#U	Meth	True	L95	U95	W	Cov	True	L95	U95	W	Cov	W	Cov	$C\pi_{01}$
.15	20/30	AGM	.022	.00	.12	.12	.99	.44	.35	.53	.18	.93	.53	.94	1.0
		MAGM	.022	.00	.12	.12	.99	.44	.34	.52	.18	.92	.52	.92	1.0
	40/15	AGM	.022	.00	.11	.11	.98	.44	.36	.51	.15	.95	.51	.94	1.0
		MAGM	.022	.00	.11	.11	.98	.44	.35	.50	.15	.94	.50	.92	1.0
	100/30	AGM	.022	.00	.07	.06	.98	.44	.40	.48	.08	.94	.48	.94	1.0
		MAGM	.022	.00	.07	.06	.98	.44	.39	.48	.08	.94	.48	.93	1.0
.50	20/30	AGM	.178	.04	.35	.31	.93	.43	.26	.57	.31	.93	.53	.93	.97
		MAGM	.178	.04	.35	.31	.93	.43	.25	.56	.31	.92	.52	.92	.97
	40/15	AGM	.178	.07	.31	.24	.95	.43	.30	.53	.23	.93	.46	.93	.98
		MAGM	.178	.07	.31	.24	.95	.43	.29	.52	.23	.93	.46	.92	.98
	100/30	AGM	.178	.11	.25	.15	.94	.43	.35	.49	.14	.95	.38	.94	1.0
		MAGM	.178	.11	.25	.15	.94	.43	.35	.49	.14	.95	.38	.94	1.0
.85	20/30	AGM	.333	.14	.55	.40	.93	.41	.20	.60	.40	.92	.45	.92	.95
		MAGM	.333	.14	.55	.40	.93	.41	.20	.59	.40	.92	.45	.92	.94
	40/15	AGM	.333	.19	.49	.30	.95	.41	.25	.54	.29	.93	.35	.93	.96
		MAGM	.333	.19	.49	.30	.95	.41	.25	.54	.29	.93	.35	.92	.95
	100/30	AGM	.333	.25	.43	.18	.95	.41	.31	.50	.18	.96	.25	.95	.98
		MAGM	.333	.25	.43	.18	.95	.41	.31	.50	.18	.96	.25	.95	.98

Legend: Table values are means over 1000 simulations, except for columns labeled 'True' values ICC= Dirichlet multinomial correlation; #C= number of clusters, #U= number of units per cluster AGM=Equation 2.2; MAGM=Equation 2.6; W=width of 95% CI= U95-L95; Cov=coverage; Simultaneous: coverage of both Lb and UB using L95 of LB, U95 of UB; $C\pi_{01}$: coverage of π_{01} using L95 of LB, U95 of UB

Fig 1. Width of 95% CI for $\pi 01$ by ICC and heterogeneity scenario
N=20 clusters, 30 units per cluster

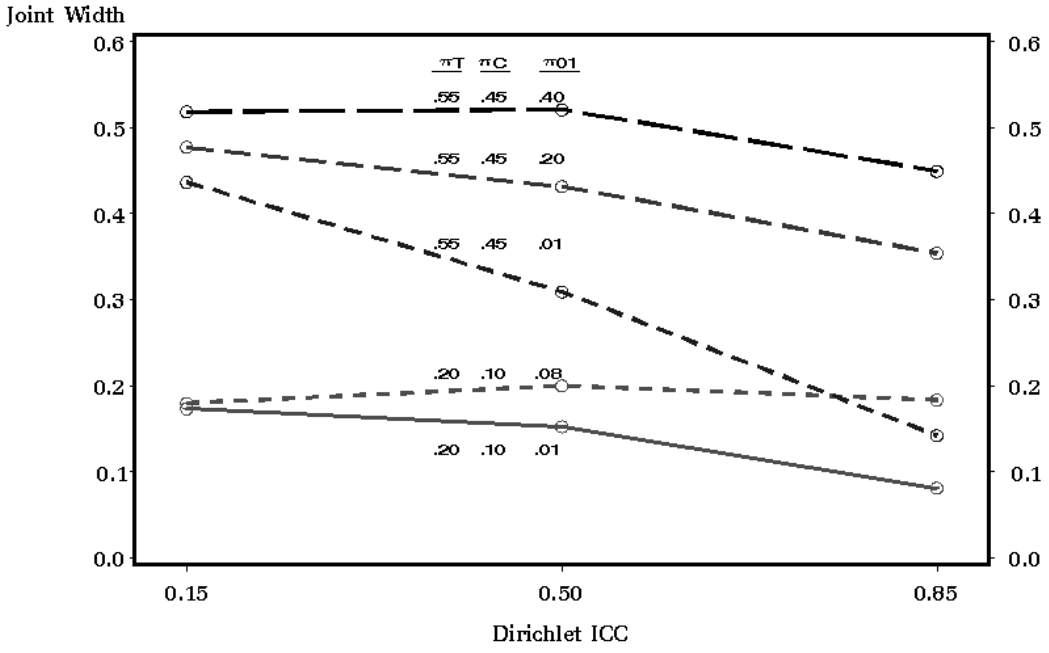
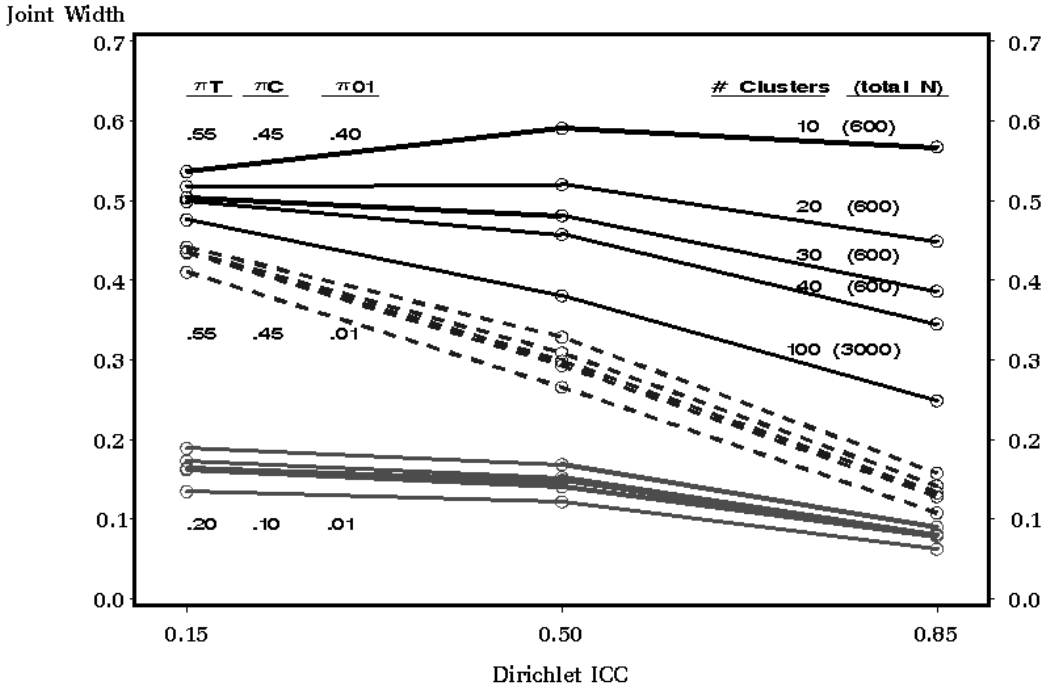


Fig 2. Width of 95% CI for $\pi 01$ by ICC, heterogeneity scenario and # clusters



intervals for the individual bounds, the simultaneous bounds and for $\pi_{01.}$.

The effect of ICC on confidence interval widths is more dramatic in the case where the marginal success probabilities are closer to 0.5. In this case, when there is high heterogeneity (π_{01}=0.4), 95% CI widths for π_{01} are reduced from around 0.5 (at ICC=0.15) to as low as 0.3 (at ICC = 0.8), and a similar reduction in width (from roughly 0.4 to 0.2) is seen in the low heterogeneity (π_{01}=0.01). This is important because CI widths of more than .20 or so are unlikely to be very useful.

Although nominal or near-nominal coverage of the true bounds was attained for most of the scenarios considered, the estimators did not give sufficient coverage of either the individual bounds or the simultaneous bounds with the combination of very high ICC and small number of clusters (20 or less) when using the fixed total sample size of 600. In results not presented, it was found that using less than 20 clusters (specifically, 10) gave very poor coverage in most scenarios. Creating a confidence interval estimator which directly takes into account the number of clusters and the ICC might greatly improve the coverage in these outlying situations.

These methods assume that the observed data consist of clusters (or blocks) that are either natural or can be created post-hoc. Post-hoc clusters can be created by first predicting the observed outcome on either T or C using all available baseline covariables, excluding treatment group, and then grouping patients by percentiles of their predicted probability of success. In order to be able to apply these methods and obtain appropriately narrow confidence intervals on bound estimators, studies would best collect data on as many baseline covariables as feasible. SAS macros will soon be available to calculate the bound estimators and confidence intervals.

Confidence intervals for the treatment risk could be used in several ways in practice. First is the case where the lower confidence limit on treatment risk is zero, and the interval width is small. Being able to conclude that the new intervention is expected to be successful for a certain proportion of the existing treatment failures, but not likely to change any of the existing treatment successes, seems ideal. But a non-zero upper bound estimate would imply that the treatment risk may be non-zero, and this may provoke interest, concern and perhaps more research. Second, if the lower estimated confidence limit was above zero, non-zero treatment risk would be concluded, and researchers would best search for patient subsets that would be better off with the standard treatment. Researchers for a new drug or treatment would likely be more satisfied with an intervention that had very low probability of failing in patients already expected or known to have success on the standard treatment.

For individual decision-making, the confidence intervals on treatment risk might be useful in some situations. An individual with no experience with either intervention might well choose the one with the largest observed marginal success, regardless of the estimated bounds on the treatment risk. On the other hand, if it was believed that the treatment risk was high, an individual with known or supposed success on the control might be hesitant to switch to an intervention with greater marginal success, even with fewer expected side effects. The gamble would be more likely if the treatment risk was thought to be low. In future work, study of the methods of using covariate information to help predict an individual's underlying category is planned.

The Dirichlet multinomial (DMN) was found to be a useful model for assessing the statistical properties of estimators for bounds on treatment effect heterogeneity because the ICC can be directly specified and because of the natural clumping of the data with higher ICC. One potential limitation of the DMN for this work is that the covariance structure is based on the underlying proportion of individuals in each category, and the corresponding structure of the intraclass between-category correlations may not be intuitive for some real situations. However, there is no reason to believe that an underlying model, allowing full specification of the covariance between the four categories of interest, would yield substantially different property assessment results. Because the parameters of interest are non-estimable (only one of two potential outcomes is observed for each unit or individual), without distributional

assumptions, at best bounds may be put on the parameters of interest.

References

Albert, J. M., Gadbury, G. L, & Mascha, E. J. (2005) Assessing treatment effect heterogeneity in clinical trials with blocked binary outcomes. *Biometrical Journal, 47*, 662-673.

Bickel, P. & Friedman, D. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics, 9*, 1196-1217.

Brier, S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika, 67*(3), 591-596.

Banerjee, T. & Paul, S. (1999). An extension of Morel-Nagaraj's finite mixture distribution for modeling multinomial clustered data. *Biometrika, 86*(3), 723-727.

Gadbury, G. L., & Iyer, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics, 56*, 882-885.

Gadbury, G. L, Iyer, H. K., & Albert, J. M. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference, 121*, 163-174.

Gadbury, G. L., Iyer, H. K., & Allison, D. (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics, 11*(4) , 313-333.

Gange, S. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician, 95*(49), 134-138.

Horowitz, J. & Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95*, 77-88.

Jensen, D. R. (1998). Multivariate distributions. In *Encyclopedia of Biostatistics*. Chichester, England: John Wiley & Sons, 2857.

Mascha, E. J. & Albert, J. M. (2006). Estimating treatment effect heterogeneity for binary outcomes via Dirichlet multinomial constraints. *Biometrical Journal* (in press).

Morel, J. & Nagaraj, N. (1993). A finite mixture distribution for modeling multinomial extra variation. *Biometrika, 80* (2), 363-71.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate B-distribution and correlation among proportions. *Biometrika, 49*, 65-82.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Essay on Principles*, Section 9. Translated in *Statistical Science* (1990) 5, 465-480.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. B. (2000). Comment on 'Causal inference without counterfactuals', by A. P. Dawid, *Journal of the American Statistical Association*, 95, 435-437.