Published in final edited form as:
Stat Biopharm Res. 2011 May 1; 3(2): 266–275. doi:10.1198/sbr.2011.10014.

Dose-Weighted Adjusted Mantel-Haenszel Tests for Numeric Scaled Strata in a Randomized Trial

Stuart A. Gansky,
University of California, San Francisco

Nancy F. Cheng, and
University of California, San Francisco

Gary G. Koch
University of North Carolina at Chapel Hill

Abstract

A recent three-arm parallel groups randomized clinical prevention trial had a protocol deviation causing participants to have fewer active doses of an in-office treatment than planned. The original statistical analysis plan stipulated a minimal assumption randomization-based extended Mantel-Haenszel (EMH) trend test of the high frequency, low frequency, and zero frequency treatment groups and a binary outcome. Thus a dose-weighted adjusted EMH (DWAEMH) test was developed with an extra set of weights corresponding to the number of active doses actually available, in the spirit of a pattern mixture model. The method can easily be implemented using standard statistical software. A set of Monte Carlo simulations using a logistic model was undertaken with (and without) actual dose-response effects through 1000 replicates for empirical power estimates (and 2100 for empirical size). Results showed size was maintained and power was improved for DWAEMH versus EMH and logistic regression Wald tests in the presence of a dose effect and treatment by dose interaction.

Keywords

chi-square; compliance; missing data; ordinal; pattern mixture model; randomization test

Introduction

Mantel-Haenszel (MH) chi-square tests for sets of 2×2 tables (e.g. stratified trials with a dichotomous treatment and dichotomous response) and their extensions for sets of $r \times c$ tables with ordinally or nominally scaled r rows and c columns have been extensively studied and utilized in randomized controlled clinical trials (e.g. Mantel and Haenszel 1959; van Elteren 1960; Koch and Edwards 1988; Stokes, Davis and Koch 2000). MH and extended MH (EMH) tests have many favorable properties, particularly being minimal assumption methods only assuming randomization. A variety of scores for ordinal rows and/or columns have been implemented in appropriate settings to use with EMH tests including integer scores, rank scores, standardized midrank scores, exponential scores, and Normal scores (e.g. Koch and Edwards 1988). However, MH and EMH tests incorporate the stratification factor as a nominally scaled factor, weighting the individual strata by their relative sample size to produce a weighted aggregate test of the independence of the rows and columns. There may be situations including randomized trials in which it would be desirable to incorporate a numerically or ordinally scaled stratification factor into the MH or EMH trend test to produce a dose-weighted adjusted MH (DWAMH) or EMH (DWAEMH) test. This paper describes a particular real world trial in which a dose-weighted adjusted test would be appropriate, proposes a DWAEMH test, provides some simulation results as well as results from the example trial, discusses limitations and advantages of these tests, and suggests some other settings, including randomized trials, in which this approach might be useful.

Motivating Example: Fluoride Varnish Randomized Controlled Clinical Trial

A recent randomized controlled clinical trial (Weintraub *et al.* 2006) of 376 preschool aged children with three parallel groups evaluated fluoride varnish (FV), which had not been adequately tested in this age group, as a potential caries preventive agent. FV is a non-invasive, inexpensive treatment that dental personnel can apply to children's teeth in-office. Children, caries-free and 6–44 months old at baseline, were randomly assigned in permuted blocks to one of three arms stratifying on the two clinics: counseling only (0 FV control), counseling plus 2 FV applications (i.e. every 12 months), and counseling plus 4 FV applications (i.e. every 6 months). Importantly, although the dentist applying FV used the same clinical set-up for the control group as the FV groups and dipped the applicator brush into water to attempt to keep parents/guardians masked to treatment group assignment, the design did not include a placebo FV group. One pediatric dentist masked to treatment group examined children 1 and 2 years post-randomization to evaluate the number of decayed or filled primary tooth surfaces (dfs). The primary outcome measure was caries incidence (dfs>0). Ultimately, a total of 280 (74%) children had at least one follow-up exam. The trial took place between October, 2000 and July, 2004. (At trial completion, the missing completely at random (MCAR) assumption was found untenable with dropouts missing at random (MAR) since there were baseline covariates relating to dropping out; more importantly for ethical reasons, children with caries at the 1-year follow-up were exited from the study as preventive intervention failures and referred for the standard of care (therapeutic FV treatment). Interested readers are referred to another publication for further information on MCAR testing and multiple imputation analyses for this trial (Gansky & Neuhaus 2009).) Due to the discrete time-to-event aspect of the study, a secondary outcome measure could be the ordinal response: no caries at either follow-up; no caries at 1 year but caries at 2 years; and caries at 1 year. Additionally, based on this design feature, a supplemental Mantel-Cox discrete time-to-event analysis could be performed.

A safety substudy of salivary fluoride content was also conducted in children in the 2FV and 4FV groups, collecting saliva samples at four times relative to FV application: pre-application, 30-minute post-application, 2-hour post-application, and 1-week post-application. Saliva samples were processed in batches to determine fluoride content. Upon analyzing these samples, the researchers discovered that between November, 2001 and August, 2002, the manufacturer had inadvertently shipped one lot of placebo FV intended for another study instead of active FV. Thus, during a 10-month period of the 46-month trial, no one received active FV for the in-office applications (Weintraub *et al.* online appendix 2006). As shown in Table 1, this resulted in only 22% of the 2FV group receiving 2 active applications and 1% of the 4FV receiving 4 active applications; 78% of the 2FV received 1 active application, while 48% of the 4FV group received 2 active applications and 31% received 3. Understandably, the investigators, sponsor, and data and safety monitoring board (DSMB) members were concerned that the protocol deviation would result in the trial no longer having adequate power to detect a clinically meaningful difference among the groups. Thus, the sample size calculation and the initial analysis plan, which called for a minimal assumption 1 degree of freedom (df) EMH correlation test to assess the relationship among the 3 ordinal treatment levels in caries incidence by 2 years, were re-evaluated and a more powerful minimal assumption method was sought.

Statistical Literature Review

Possible existing statistical methods, particularly those for non-compliance/non-adherence were surveyed for applicability in this unusual scenario. An instrumental variables approach (e.g. Greenland, 2000) was considered as a potential solution, but did not seem appropriate as very few participants were treated as planned; even with a sufficient number of participants with the treatment as planned, it is unclear if this method would have been appropriate. Causal effects modeling with potential outcomes (e.g. Little & Rubin, 2000) was evaluated, but since the source of non-compliance was temporal in nature and external to the study (i.e. active treatment depended only on enrollment date and not through the will of participants) groups such as defiers, never takers, and always takers did not exist. Finally, the pattern mixture model (PMM) for randomization-based MH test method (Sato, 2001; Matsuyama, 2002) was considered relevant for this trial.

To apply the PMM for MH, stratify participants on the eligible dosing; i.e., the number of active applications they *would* have received in this trial if they were all assigned to the 4FV (every 6 month) group as in the schematic in Figure 1. This can be determined based on accrual date and knowledge of usage of the active and inactive lots. Of the 2^4=16 possible patterns of active/inactive across the 4 treatment periods (baseline, 6 months, 12 months, and 18 months), only 9 were actually realized: AAAA, AAAI, AAIA, AIAA, IAAA, AAII, AIIA, IIAA, and AIII, where A=active and I=inactive application (or, for purposes of illustrating the DWAEMH method, missed visit). Stratifying on these patterns would be akin to the methods of Sato (2001) and Matsuyama (2002).

However, since the stratification variable is the eligible dosing (number of possible active applications), that factor has a natural metric on the integer scale that could be used for a dose-response type analysis (i.e. 3 for AAAI, AAIA, AIAA, and IAAA; 2 for AAII, AIIA, and IIAA; and 1 for AIII; AAAA was categorized with those receiving 3 active since there was only one participant in that group). Thus, in addition to the usual row and column weights of EMH tests, adding stratum weights could account for eligible dose (possible number of active applications).

Methods

The standard EMH chi-square statistic (Q_{EMH}) across the H strata is calculated as

$$Q_{EMH}=\left(\sum_{h=1}^{H}(\mathbf{n}_h-\mathbf{m}_h)'\mathbf{A}_h'\right)\left(\sum_{h=1}^{H}\mathbf{A}_h\mathbf{V}_h\mathbf{A}_h'\right)^{-1}\left(\sum_{h=1}^{H}\mathbf{A}_h(\mathbf{n}_h-\mathbf{m}_h)\right)$$

where $\mathbf{n}_{\boldsymbol{h}}$ is the observed count vector with elements n_{hij} defined in Figure 1, $\mathbf{m}_h = n_h(\mathbf{p}_{h\bullet}{}^* \otimes \mathbf{p}_h{}^*{}_\bullet$ is the expected count vector, $\mathbf{p}_h{}^*{}_\bullet$ and $\mathbf{p}_{h\bullet}{}^*$ are the row and column marginal proportion vectors, ⊗ is the left Kronecker product operator (which multiplies the matrix on the left by the elements of the matrix on the right),

$\mathbf{V}_h=n_h^2(\mathbf{D}_{\mathbf{p}_{h\bullet *}}-\mathbf{p}_{h\bullet *}\mathbf{p}_{h\bullet *}')\otimes(\mathbf{D}_{\mathbf{p}_{h^*\bullet}}-\mathbf{p}_{h^*\bullet}\mathbf{p}_{h^*\bullet}')/(n_h-1)$ is the covariance matrix for the h-th stratum, $\mathbf{D}_{\mathbf{p}_h}$ is a diagonal matrix with elements of vector $\mathbf{p}_h$ on the main diagonal, and $\mathbf{A}_{\boldsymbol{h}}$ is a non-redundant linear operator matrix of row and column scores: $\mathbf{A}_{\boldsymbol{h}}= \mathbf{c}'_h \otimes \mathbf{r}'_h$, where $\mathbf{r}_{\boldsymbol{h}}$ is the vector of row scores and $\mathbf{c}_{\boldsymbol{h}}$ is the vector of column scores. The originally planned standard application of Q_{EMH} to the fluoride varnish trial would define two strata (H=2) for the clinics. Although Q_{EMH} does not require homogeneity of odds ratio across strata for its use, it can have reduced power with heterogeneous odds ratios, particulary when stratumspecific effects are in opposite directions that tend to cancel each other (i.e., a certain type of stratum by row (treatment) interaction).

The dose-weighted adjusted EMH (DWAEMH) test (Q_{DWAEMH}) would extend the standard EMH test with a set of stratum scores, w_h, augmenting the usual score matrix for a new linear operator matrix defined as $\mathbf{A}^{*}_{\boldsymbol{h}}=\mathrm{w}_h\mathbf{A}_h=\mathrm{w}_h\mathbf{c}'_{\boldsymbol{h}}\otimes\mathbf{r}'_{\boldsymbol{h}}$. Then Q_{DWAEMH} is a modification of the EMH correlation statistic (Q_{CSMH}) with stratum scores (w_h) added (Stokes, Davis and Koch, 2000) which can be shown as:

$$Q_{CSMH}=\frac{\left[\sum_{h=1}^{H} w_h \sum_{i=1}^{r}\sum_{j=1}^{c}\left(r_{hi}-\bar{r}_h\right)\left(c_{hj}-\bar{c}_h\right)n_{hij}\right]^2}{\sum_{h=1}^{H} w_h^2\left[\sum_{i=1}^{r}\left(r_{hi}-\bar{r}_h\right)^2 n_{hi\bullet}\sum_{j=1}^{c}\left(c_{hj}-\bar{c}_h\right)^2 n_{h\bullet j}\right]\left(n_h-1\right)^{-1}}$$

where $\bar{r}_h$ and $\bar{c}_h$ are the mean row and column scores in the $\underline{h}$-th stratum, respectively, and is similar in spirit to a stratum-weighted Pearson correlation coefficient between rows and columns.

Generally with a 3-level ordinal treatment and binary response, the usual EMH test would have score vectors $\mathbf{r}_h$ = [0 1 2]′ and $w_h\mathbf{c}_h$ = [0 1]′ for all h=1, 2, … H strata, while the DWAEMH test would have score vectors $\mathbf{r}_h$ = [0 1 2]′ and $w_h\mathbf{c}_h$ = [0 w_h]′ for stratum weights w_h which could involve integer, standardized midrank (modified ridit), exponential, Normal, or other scores. In this particular example trial with binary response, the usual EMH test would have score vectors $\mathbf{r}_h$ = [0 1 2]′ and $w_h\mathbf{c}_h$ = [0 1]′ for all h=1, 2, … H (= 2 clinics × 3 eligible dosing groups = 6) strata, while the DWAEMH test would have score vectors $\mathbf{r}_h$ = [0 1 2]′ and $w_h\mathbf{c}_h$ = [0 w_h]′ where w_h=1, 2, 3 according to the eligible dosing (i.e. number of active applications which was 1, 2 or 3) within each clinic, as a natural choice. More generally, Q_{CSMH} can have stratum scores $w_h\mathbf{c}_h$ varying across strata according to categories of one or more stratification factors (such as eligible dosing groups), while remaining constant across strata for the other stratification factor(s) (such as clinics) in the overall cross-classification of these factors (H=6 for 2 clinics × 3 eligible dosing groups).

Simulations

A logistic model was used to simulate power for 1000 replicates under various conditions, each with a sample size of 189 (63 per treatment group each with 21 at each of 3 doses), which was the projected retained sample size for the motivating trial (but with a null center effect) at the time of the power analysis for the DSMB meeting. The simulated logistic probability (p) for the conditions was

$$p=exp\{\alpha-\beta t+\gamma(2-w)+\phi t(2-w)\}/[1+exp\{\alpha-\beta t+\gamma(2-w)+\phi t(2-w)\}]$$

or, equivalently,

$$logit(p)=\alpha-\beta t+\gamma(2-w)+\phi t(2-w),$$

where the intercept α=0; the treatment effect β=0.1 to 0.7 in 0.1 increments; the treatment group for number of annual applications t=0,1,2 (corresponding to 0FV, 2FV, and 4FV, respectively); the dose effect γ=0, 0.25, 0.50, 0.75; the number of active applications (eligible dosing) covariate w=1,2,3; and the treatment by dose interaction Φ=0, 0.25, 0.50.

Six combinations of (γ, Φ) were used to assess power: (0.25, 0), (0.50, 0), (0.50, 0.25), (0.75, 0.25), (0.25, 0.25), and (0.50, 0.50). To assess size (Type I error), α=β=γ=Φ=0 was used with 2100 replicates to provide an appropriately small standard error (yielding confidence interval half-widths less than 0.01). For both power and size, Bernoulli random variables were generated from the logistic probability (p) with the SAS call ranbin function with attention to maintain unique seed values. From the prior equation for *logit(p)*, the contrast for the difference between t=0 and t=2 is

$$\{\alpha-2\beta+\gamma(2-w)+2\phi(2-w)\}-\{\alpha+\gamma(2-w)\}=-2\beta+2\phi(2-w),$$

so for w=1, 2, 3, the equal weighted treatment effect (odds ratio) is estimated by $\{(-2\beta+2\Phi) + (-2\beta) + (-2\beta-2\Phi)\}/3 = -2\beta$, while the dose-weighted treatment effect with weights 1, 2, 3 is estimated by $\{(-2\beta+2\Phi)/6 + (-4\beta)/6 + (-6\beta-6\Phi)/6 = -2\beta-2/3\Phi$; for the contrast of the difference between t=0 and t=1, the equal weighted treatment effect is estimated by $-\beta$, while the dose-weighted treatment effect is estimated by $-\beta-1/3\Phi$.

For the simulated null scenario (zero effects), Q_{EMH} and Q_{DWAEMH} were determined and the proportion of their corresponding p-values less than or equal to 0.05 yielded empirical Type I error estimates. Empirical Type I error estimates were determined for Wald test statistics (Q_W) from a standard main effects logistic regression model with an intercept and treatment effect (which is misspecified when $\Phi \neq 0$), which was used to estimate an equal weighted treatment effect ($-\beta$), and a standard logistic regression model with intercept, treatment effect, dose effect and treatment by dose interaction which was used to estimate a dose-weighted treatment effect ($-\beta -\Phi/3$). In addition, empirical Type I error estimates were determined for Wald test statistics (Q_W) from standard conditional logistic regression (CLR) models with eligible dose as a stratification variable. For the other 42 combined scenarios with varying β, γ, and Φ, empirical power was determined similarly.

Results

Simulations

Results of the simulations are shown in panel plots in Figures 2–4 with the proportion of p-values less than or equal to 0.05 on the vertical axis and the effect size (β) on the horizontal axis. Empirical Type I error (size), which corresponds to effect size of zero on the horizontal axis, is shown in each graph; by definition there is no dose effect for this scenario so the dose effects start at effect size of 0.1. All the methods show correct size (Type I error) with EMH 0.043, DWAEMH 0.049, Wald tests from logistic models 0.037 in the misspecified model, 0.037 in the equal weighted treatment model and 0.047 in the dose weighted interaction model, and Wald tests from CLR 0.037 in the equal weighted treatment model and 0.041 in the dose weighted interaction model; all Type I error estimates had standard errors of 0.0041 to 0.0047. Figures 2–4 illustrate the results with all 6 panels graphing the same Type I error (size) results and the same no dose effect power results as compared to power for the different dose and treatment by dose combinations. In Figures 2–4, the no dose effect (i.e. α=γ=Φ=0) power results are shown as dashed lines with open circle symbols, the dose effect ($\gamma \neq 0$, with or without treatment by dose interaction) power results are shown as solid lines with closed circle symbols, the dose effect with dose weight ($\gamma \neq 0$, with or without treatment by dose interaction) power results are shown as dotted lines with open circles, EMH results are shown in red, DWAEMH results shown in blue, Wald results shown in green, and conditional logistic regression results shown in gold. In Figure 2 showing (γ, Φ) values of (0.25, 0) and (0.50, 0), the Wald tests and EMH power curves are very similar when a dose effect with no interaction exists and when no dose effect exists;

while the blue DWAEMH curves (with or without a dose effect) show lower power. In Figure 3 with (γ, Φ) values of (0.50, 0.25) and (0.75, 0.25), DWAEMH with a treatment by dose effect (solid blue line) and EMH with no dose effect (dashed red line) show the highest power. Figure 4 with (γ, Φ) values of (0.25, 0.25) and (0.50, 0.50), shows a gap between the curve for DWAEMH with a treatment by dose effect (solid blue line) and the Wald (solid green line) and EMH curves with dose effect (solid red line); the Wald with a dose effect (short-dashed green line) and Wald from CLR (short-dashed gold- line) with a dose effect had lower power. These results show the weighted methods have increased power as the treatment by dose interaction (Φ) increases.

Example Trial

At the end of the trial, 278 participants who would have been assigned to 1, 2, or 3 active FV applications if assigned to the 4FV group had evaluable follow-up data with dfs counts ranging from 0 to 15. (Two participants who would have been assigned to 0 active FV were excluded from these analyses.) The 0FV group had 42% with caries by two years, the 2FV group had 25%, and the 4FV group had 16% 2 year caries incidence; more detail is shown in Table 2.

Table 3 displays the results of nonparametric tests. Based on center strata, with binary dfs>0 (incidence), Q_{EMH}=14.5, p=0.0001, while based on the 6 cross-classified strata of center and dose strata with dose weights Q_{DWAEMH}=14.4, p=0.0002. For categorized dfs (72% 0, 11% 1, 10% 2–4, 8% 5–15), center-stratified integer score nonparametric Q_{EMH}=13.1, p=0.0003, while center by dose strata with dose weights Q_{DWAEMH}=12.4, p=0.0004. Additionally, for the 3-level ordinal outcome of no caries by 2 years, caries at 2 years and caries at 1 year with integer scores and center stratification (data shown in Table 2) had Q_{EMH}=11.9, p=0.0006, while center by dose strata with dose weights Q_{DWAEMH}=11.34, p=0.0008.

Center by dose interaction was not statistically significant. For comparison (Table 4), from standard logistic regression with only an indicator for center and a linear treatment effect (0,1,2), the 1 df equal weight Wald test Q_W=14.0, p=0.0002, while from the model with interaction the dose-weighted Wald test Q_W=12.8, p=0.0004. The corresponding standard conditional logistic regression model, with a linear treatment effect and linear treatment by dose effect stratifying on the 6 cross-classified strata of center and the eligible dosing (number of active FV applications if assigned to the 4FV group), showed the interaction to be non-significant (p=0.5488); dropping the interaction produced a 1 df conditional equal weight contrast Wald test of 14.0 with p=0.0002; the corresponding dose weighted contrast from standard conditional logistic regression (which included the interaction) had a Wald test of 12.3 with p=0.0004. The equal weight contrast from standard conditional logistic regression had an odds ratio of 0.52 (exact 95% confidence interval of 0.36–0.75). This odds ratio indicated that one planned FV application per year reduced the odds of caries incidence by about half and two planned FV applications per year reduced the odds of caries incidence by almost three-quarters. Additionally, the Mantel-Cox test of the life table approach of the data in Table 2 resulted in $Q_{EMH\text{-}MC}$ =14.4, p=0.0001, while a dose-weighted adjusted Mantel-Cox test stratifying each 1 year follow-up period on whether 1 or 2 active applications would have been received if assigned to the every 6 months group resulted in $Q_{DWAEMH\text{-}MC}$=12.3, p=0.0005.

Discussion

The proposed dose adjusted extended Mantel-Haenszel test (DWAEMH) shows promise as a way to incorporate numeric or ordinal stratification factors into a nonparametric analysis. DWAEMH, like EMH and logistic regression, provides proper size but can yield additional power when there is a treatment by dose interaction. In confirmatory protocols, particularly those in regulatory environments, the statistical analysis plan should include DWAEMH as the *a priori* method, or prior to beginning analysis an amendment should be filed.

The proposed DWAEMH approach has all the same limitations as EMH tests. Namely, it assumes: row and column weights have good external justification and are chosen *a priori*; row sample sizes are adequate for asymptotic properties to apply (e.g. $n_{hi\bullet} \geq 5$); and of course that the treatment is assigned by a proper implementation of a valid randomization schedule (or that the data arise from a process like stratified simple random sampling). The other assumptions should be readily met in well conducted randomized trials. DWAEMH also assumes there is good external justification for, and *a* priori specification of, stratum weights. Weights based on a function of dose, such as log of dose, square root of dose, or cube root of dose, would be readily justifiable; as the root increases, the weights approach equal weighting. In this particular trial, the DWAEMH analysis assumes no order effect of the active applications; e.g. that AAII is no different in incidence than IIAA. The DWAEMH analysis weights the data based on dose, which may lead the reader to interpret this as a shift from an intention-to-treat (ITT) approach toward a protocol-compliant or as-treated approach. However, it is important to realize that the weights are based on the eligible dosing (the number of active FV applications the participant would have received if randomized to the 4FV group) and not the actual doses received (based on eligible dosing and randomization), so that a separate statistic is estimated in each mutually exclusive eligible dosing stratum and then aggregated across strata which is still in the spirit of an ITT approach. To apply this method as an ITT approach in a regulatory setting, an amendment to the statistical analysis plan would have to be filed before the data could be analyzed. It would be possible to use preliminary data from other studies to perform simulations as in this paper to characterize the sort of power gains that might be achieved without increasing Type I error.

This stratum score approach could be utilized for Mantel-Cox time-to-event analyses in which the strata are the usual risk sets at each time cross-classified with the dose. To implement that approach in this example trial, the number of active doses was modified from 1 through 3 as shown in Table 2, to 1 or 2 in each annual risk set (zeros were dropped since they are not informative). For example, the AIII group only contributed to the 1-year follow-up, while the IIAA group only contributed to the 2-year follow-up.

Although we expected the motivating example trial would illustrate the advantages of this method, for this example the dose-weighted method did not show an advantage; however, it did not show a loss in the ability to detect the effect either. This is coherent with the simulations when there is a modest treatment by dosing interaction effect. In addition, the example had circumstances different than usual non-adherence; instead the participants had no conscious role in not receiving the assigned number of treatments. While unusual, there are other circumstances which could result in a similar situation, such as a natural disaster, break in funding, temporary institutional review board shutdown, clinic closure, or extended staff illness, injury or turnover, which could interrupt study applications. For example, in a recent trial in San Diego County by this same research group, wildfires interfered with study procedures for a number of weeks. Other situations in which this DWAEMH approach might be useful for baseline ordinal or numeric strata would include disease severity or stage, disease correlate status (e.g. fluoride level, tobacco use, age), self-efficacy score or internal health locus of control score, or socioeconomic status.

Finally, this paper exemplifies where a real life example motivated modifying an analytic method. Simulations helped illustrate the properties of the method under different scenarios to confirm it has proper Type I error and to delineate when it can perform better than traditional methods. Interestingly when the motivating trial was completed, these data did not ultimately benefit from the modified method; although treatment by dosing effect trended in the anticipated direction, enabling the weighted odds ratio estimate to indicate a somewhat stronger treatment benefit than the unweighted estimate, the variance (on the log odds scale) also increased, leading to somewhat less sensitivity to detect the effect per FV application.

Acknowledgments

The authors thank Drs Jane Weintraub, Sara Shain, Francisco Ramos-Gomez, Bonnie Jue, Charles Hoover, John Featherstone, and colleagues for conducting the fluoride varnish randomized clinical trial, the study's data and safety monitoring board members for helpful suggestions and guidance with trial, and finally the trial participants. We thank Michael Hussey and Jill El-Khorazaty for their suggestions and two anonymous reviewers for their helpful comments in improving the manuscript. Any remaining errors or ambiguities are our own. In addition, the authors express their appreciation to Dr Lisa Lavange for coordinating this special issue and Dr John Preisser for organizing the Festschrift in honor of Dr Gary Koch, for guest editing this special issue, and for providing helpful feedback to clarify the paper.

This research was supported in part by the following grants and cooperative agreements: US DHHS/NIH/NIDCR P60DE013058, NIDCR & NCMHD U54DE014251, and NIDCR U54DE019285.

References

Gansky, SA.; Neuhaus, JM. Missing Data and Informative Cluster Sizes. In: Lesaffre, E.; Feine, J.; Leroux, B.; Declerck, D., editors. Statistical and Methodological Aspects of Oral Health Research. New York: Wiley; 2009.

Greenland S. An Introduction to Instrumental Variables for Epidemiologists. International Journal of Epidemiology. 2000; 29:722–729. [PubMed: 10922351]

Koch, GG.; Edward, S. Clinical Efficacy Trials with Categorical Data. In: Peace, KE., editor. Biopharmaceutical Statistics for Drug Development. New York: Marcel-Dekker; 1988. p. 403-457.

Little RJ, Rubin DB. Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches. Annual Review of Public Health. 2000; 21:121–45.

Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute. 1959; 22:719–48. [PubMed: 13655060]

Matsuyama Y. Correcting for Non-compliance of Repeated Binary Outcomes in Randomized Clinical Trials: Randomized Analysis Approach. Statistics in Medicine. 2002; 21:675–87. [PubMed: 11870809]

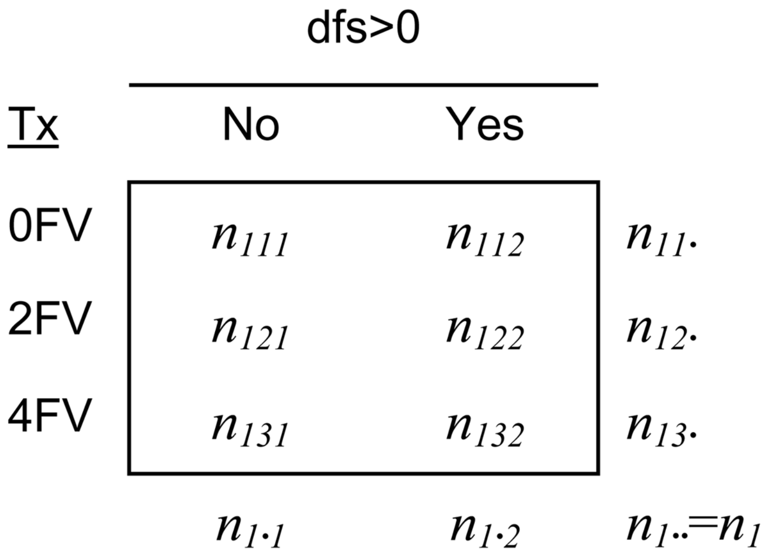
Ramos-Gomez, Chung LH, Beristain R, Santo W, Jue B, Weintraub JA, Gansky SA. Recruiting and Retaining Pregnant Women from a Community Health Center at the U.S.-Mexico Border for the Mothers and Youth Access (MAYA) Clinical Trial. Clinical Trials. 2008; 5:336–46. [PubMed: 18697848]

Sato T. A Method for the Analysis of Repeated Binary Outcomes in Randomized Clinical Trials with Non-compliance. Statistics in Medicine. 2001; 20:2761–74. [PubMed: 11523081]

Stokes, ME.; Davis, CS.; Koch, GG. Categorical Data Analysis using the SASReg; System. 2. Cary, NC: SAS Institute, Inc; 2000.

van Elteren PH. On the Combination of Independent Two-sample Tests of Wilcoxon. Bulletin of the International Statistical Institute. 1960; 37:351–361.

Weintraub, JA.; Ramos-Gomez, F.; Jue, B.; Shain, S.; Hoover, CI.; Featherstone, JDB.; Gansky, SA. Fluoride Varnish Efficacy in Preventing Early Childhood Caries; Journal of Dental Research. 2006. p. 172-6.with online Data Supplement http://jdr.sagepub.com/cgi/data/85/2/172/DC1/1



	dfs>0		
Tx	No	Yes	
0FV	n_{111}	n_{112}	$n_{11\cdot}$
2FV	n_{121}	n_{122}	$n_{12\cdot}$
4FV	n_{131}	n_{132}	$n_{13\cdot}$
	$n_{1\cdot 1}$	$n_{1\cdot 2}$	$n_{1\cdot\cdot}=n_1$

Figure 1.

Schematic of one stratum

(*h*=1 Clinic 1 with eligible dosing pattern AAAI)

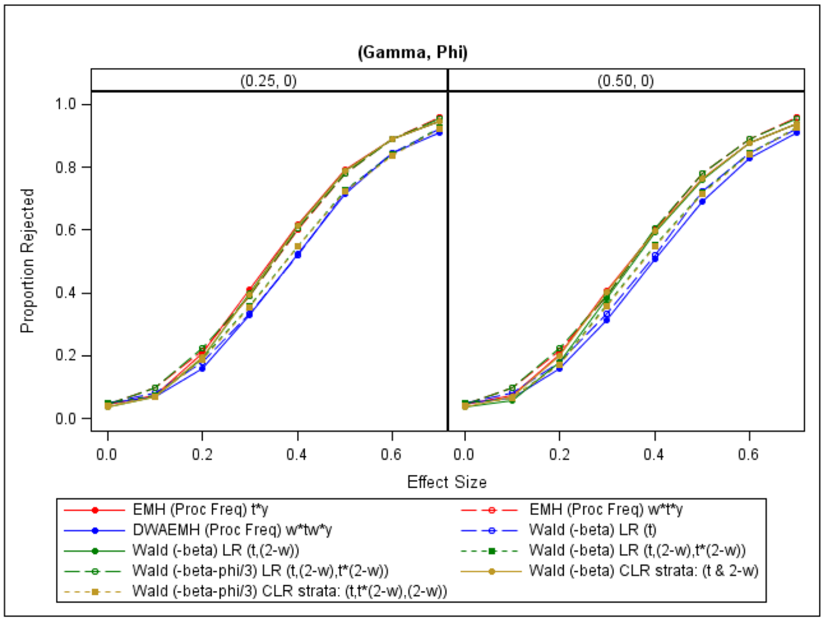


Figure 2.

Empirical Power and Size for Extended Mantel-Haenszel (EMH), Dose Adjusted EMH, Wald, and Conditional Logistic Regression (CLR) Wald Methods with No Treatment by Eligible Dosing Interaction (*Φ*=0)

No Dose Effect is for the scenario with $\gamma = \Phi = 0$.

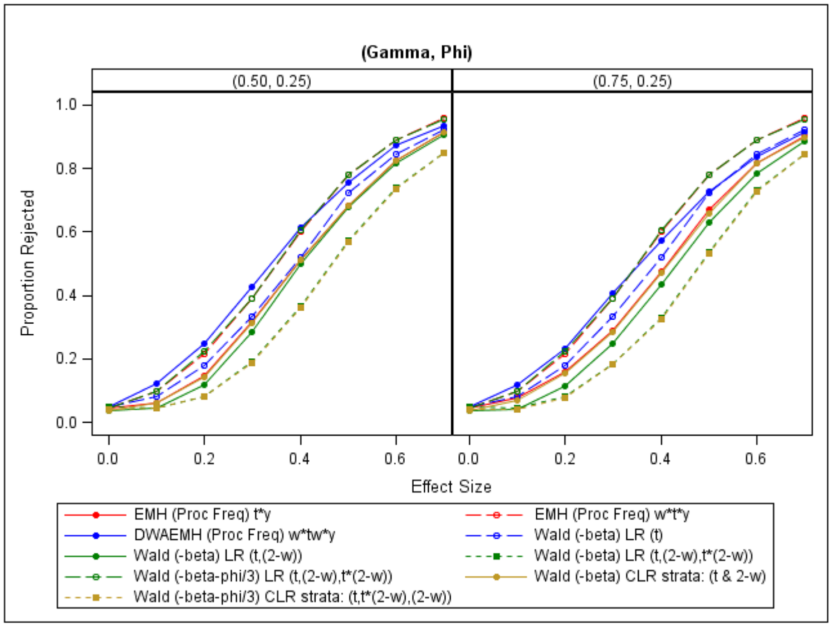


Figure 3.

Empirical Power and Size for Extended Mantel-Haenszel (EMH), Dose Adjusted EMH, Wald, and Conditional Logistic Regression (CLR) Wald Methods with Treatment by Eligible Dosing Interaction (Φ=0.25)

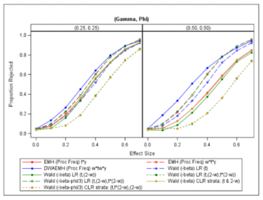


Figure 4.

Empirical Power and Size for Extended Mantel-Haenszel (EMH), Dose Adjusted EMH, Wald, and Conditional Logistic Regression (CLR) Wald Methods with Simple Treatment Effect Equal to Treatment by Eligible Dosing Interaction ($\gamma = \Phi$)

Table 1

Planned versus active applications received in example trial (N=280)

# Active Applications	2 Planned (%)	4 Planned (%)
0	0	1
1	78	19
2	22	48
3	0	31
4	0	1

Table 2

Counts of ordinal time to caries Incidence at the 2 Follow-up Visits by Treatment Group, Eligible Dosing and Center

Center	Eligible Dosing	FV Tx Group	Caries at 1 Year	Caries at 2 Years	No Caries at 2 Years	Total	% No Caries at 2 Years
1	1	0	0	1	3	4	75
		2	0	2	3	5	60
		4	0	0	3	3	100
	2	0	8	2	8	18	44
		2	1	2	14	17	82
		4	0	1	17	18	94
	3	0	7	5	18	30	60
		2	8	2	16	26	62
		4	3	0	20	23	87
2	1	0	0	0	3	3	100
		2	0	0	2	2	100
		4	1	0	2	3	67
	2	0	7	2	14	23	61
		2	1	1	14	16	88
		4	5	1	13	19	69
	3	0	5	4	12	21	57
		2	3	3	21	27	78
		4	2	1	17	20	85
Total			51	27	200	278	72

Eligible Dosing is the number of active FV applications the participant would have received if randomized to the 4FV group.

Table 3

Caries Incidence/Increment Chi-Square Tests

Method	Cross-classification	Q	p-Value
Extended Mantel-Haenszel	center*tx*incidence	14.52	0.0001
	center*dose*tx*incidence	14.16	0.0002
	center*tx*increment category	13.07	0.0003
	center*dose*tx*increment category	12.79	0.0003
	center*tx*increment	7.88	0.0050
	center*dose*tx*increment	7.63	0.0057
	center*tx*ordinal caries	11.93	0.0006
	center*dose*tx*ordinal caries	11.54	0.0007
Mantel-Cox (MC)	center*year*tx*incidence	14.42	0.0001
Dose Weight Adjusted EMH	center*dose*tx*(dose*incidence)	14.37	0.0002
	center*dose*tx*(dose*increment cat)	12.40	0.0004
	center*dose*tx*(dose*increment)	9.13	0.0025
	center*dose*tx*(dose*ordinal caries)	11.34	0.0008
Dose Weight Adjusted MC	center*dose*year*tx*incidence	12.29	0.0005

Q = chi-square; tx = treatment; incidence=dfs>0; increment=dfs; increment category= dfs category of 0, 1, 2–4, 5–15; ordinal caries=caries at 1 year, caries at 2 years, no caries at 2 years

Table 4

Models of Caries Incidence by the Second Follow-up Visit

Regression Method	Model	Parameter	Estimate	SE	Chi-square	p-Value	Odds Ratio	Conf Int
Logistic	center, t	*β*	−0.6604	0.1764	14.01	0.0002	0.52	0.37–0.73
	center, t, (2-w)	*β*	−0.6617	0.1768	14.01	0.0002	0.52	0.36–0.73
		γ	−0.1591	0.2227	0.51	0.4748	0.85	0.55–1.32
	center, t, (2-w), (2-w)t	*β*	−0.5880	0.2176	7.30	0.0069	0.56	0.36–0.85
		γ	−0.2829	0.3140	0.81	0.3675	0.75	0.41–1.39
		Φ	0.1617	0.2856	0.32	0.5713	1.18	0.67–2.06
	dose weighted	*β* +*Φ*/3	−0.6419	0.1796	12.76	0.0004	0.53	0.37–0.75
Conditional Logistic (strata: ctr × (2-w))	t	*β*	−0.6454	0.1752	13.58	0.0002	0.52	0.37–0.74
	t, (2-w)t	*β*	−0.5668	0.2166	6.85	0.0089	0.57	0.37–0.87
		Φ	0.1728	0.2883	0.36	0.5488	1.19	0.68–2.09
	dose weighted	*β* +*Φ*/3	−0.6244	0.1779	12.32	0.0004	0.54	0.38–0.76
Exact Conditional (strata: ctr × (2-w))	t	*β*	−0.6454	N/A	N/A	0.0002	0.52	0.36, 0.75
	t, (2-w)t	*β*	−0.5632	N/A	N/A	0.0101	0.57	0.36–0.88
		Φ	0.1721	N/A	N/A	0.6433	1.19	0.64–2.17

t = treatment

w = eligible dosing

dose weighted is estimated from *β*+*Φ*/3 for *t*=1 vs *t*=0 for comparability to the models with main effects for treatment