The Impact of (Not) Stratifying Analyses by Site When Randomization Was Stratified

John W. Seaman III* Brian L. Wiens†

Abstract

We consider analyses of clinical trials of novel medical devices when the randomized assignment of treatment to subject was stratified by investigative site. In pharmaceutical studies, it is common to stratify the analysis by the same factor(s) used to stratify the randomization. For medical devices, advice from regulators has often been to report an unstratified analysis as primary. To evaluate the impact, we report simulations of a clinical trial with a dichotomous primary endpoint and a non-inferiority hypothesis. When the outcome differs by site (due to differing surgeon skill, heterogeneous study subjects or other differences) the stratified analysis produced superior control of the type I error rate. Power was maintained or improved. When outcome did not differ by site, the stratified analysis maintained the type I error rate with minor loss of power. Assessment of consistency of results among sites is important to properly interpret the stratified analysis. We conclude that the stratified analysis has few drawbacks while providing important advantages in both size and power.

Key Words: Non-inferiority, Clinical Trials

1. Background

Suppose one wants to compare an investigational device to the current standard of care. The primary endpoint of interest is a binary outcome of treatment success. Suppose further that the primary analysis is one of non-inferiority, to demonstrate that the success rate with the investigational device is not markedly worse than the success rate with the standard of care. After discussion with clinical colleagues, we set the non-inferiority margin to 10 percentage points. Finally, suppose the study will enroll at several sites. The question then becomes whether to analyze the data with a stratified analysis or an unstratified analysis.

For the stratified analysis, the hypotheses tested are:

$$H_0 : \pi_{D_i} - \pi_{S_i} \leq -0.10 \text{ for all i}$$
$$H_1 : \pi_{D_i} - \pi_{S_i} > -0.10 \text{ for at least one i}$$

where π_{Ki} is the success rate with method K (K=D for the investigational device; K=S for the standard of care) at site i (i = 1, ..., n).

For the unstratified analysis, the hypotheses tested are:

$$H_0 : \pi_D - \pi_S \leq -0.10$$
$$H_1 : \pi_D - \pi_S > -0.10$$

where π_K is the success rate with method K (K=D for the investigational device; K=S for the standard of care).

*Alcon Laboratories, 6201 South Freeway, Fort Worth, Texas 76134

†Portola Pharmaceuticals, 270 East Grand Avenue, South San Francisco, CA 94080

Note that the null hypothesis for the stratified analysis implies the null hypothesis for the unstratified analysis. However, the converse is not true: it is possible that the null hypothesis for the stratified analysis is false but the null hypothesis for the unstratified analysis is true.

The analysis we will investigate for the stratified case is the method developed by Yanagawa, Tango and Hiejima (1994) while the analysis chosen for the unstratified case is the method developed by Farrington and Manning (1990). These tests will be referred to as YTH and FM, respectively, in the remainder of the document. In both cases, a restricted maximum likelihood estimate is calculated (under the restriction of equality in the null hypothesis), and the estimate is used in calculating a test statistic that is asymptotically normal. A one-sided test, $\alpha = 0.025$, will be used with either method.

With either method, a confidence interval will be reported that will include all values of δ for which the null hypothesis is not rejected: $\{\delta : H_0 : \pi_{D_i} - \pi_{S_i} \leq \delta \text{ is not rejected}\}$ for the stratified YTH analysis or $\{\delta : H_0 : \pi_D - \pi_S \leq \delta \text{ is not rejected}\}$ for the unstratified FM analysis. (See Sterne, 1954, for a discussion of these confidence intervals.) Thus, the size of the test directly corresponds to the coverage probability of the confidence interval. Additionally, the confidence interval with the YTH method can be calculated under the assumption that the difference in success rates is identical in each stratum, even if the underlying success rates vary by stratum. Thus, the interval can be used to describe a common difference in success rates, taking stratification into account.

2. Theoretical discussion of stratification

In this section, we will briefly discuss some of the reasons for stratifying (or not stratifying) analyses. This is not intended to be a comprehensive literature review, but a brief discussion of some commonly discussed points of view.

It is common to interpret R.A. Fisher's thoughts on randomization thusly: As ye randomize, so shall ye analyze (Senn, 2004). That is, if a clinical trial is randomized using a stratification factor, the analysis should use the same stratification factor as a covariate in the model. The purpose is two-fold: to appeal to the randomization as the only basis of inference (Fisher, 1956), and to increase precision of the resulting estimator.

Pocock et al. (2002) discussed three statistical goals of any analysis: to achieve an appropriate p-value, to achieve an unbiased point estimate and a confidence interval with appropriate coverage, and to improve the power of the trial to detect real differences. Pocock et al. further discussed whether to automatically stratify an analysis for any factors used to stratify the randomization, and whether to automatically stratify an analysis for centers in a multicenter trial. (Pocock et al., 2002, page 2925) The advice was ambiguous, recommending that stratification factors that have small correlation with outcomes should not be used as covariates in a model, recommending that ambiguities in handling centers with small sample sizes should be resolved, and noting that, in the collective experience of the authors, it "rarely" makes a large difference.

More recently, Kahan and Morris (2011) noted that ignoring stratification results in "confidence intervals that are too wide, type I error rates that are too low and a reduction in power." In particular, variables that are expected to be related to the outcome should be included as covariates. Finally, they caution that consequences of unadjusted analyses in

stratified clinical trials are not fully appreciated.

In a discussion by Nicholas (2014), investigating analyses unadjusted for stratification factors found that there was a decrease in power for superiority testing and for analyses of dichotomous outcomes. For continuous outcomes, it was noted that there was both a decrease in power and a decrease in type I error. Finally, for non-inferiority, the investigation pointed to an increase in power and type I error.

This small set of references is, in our experience, typical of the discussion in the literature. There is not complete consensus over whether all covariates used for randomization should be included as covariates in an analysis, but agreement that covariates that relate strongly to outcomes should be included. With very limited prior information on the site-to-site variability with the investigational device, it is not known with certainty whether the site is strongly related to the outcome (successful treatment).

3. Simulations

To understand the performance of the proposed analysis methods, a small simulation study was run. Code was generated to calculate the test statistics of the YTH test and the FM test. The code for the YTH method was validated by comparing output from the code to an example in the paper by Yanagawa, Tango and Hiejima. The code for the FM test was validated by comparing to an example run in SAS, in which the FREQ procedure has an option to report the results of the FM test. Additionally, both were programmed by two independent programmers and the results compared to ensure matching output.

In every situation, the simulations used six sites. Sample size per site was either 50/site at all six sites or 40/site at three sites and 60/site at three sites, corresponding to a balanced and an unbalanced enrollment under six sites, and 300 subjects total. Success rates of 1% to 99% were assumed in the simulations. For each situation, 10,000 simulations were run. For simulating a true type I error rate of 2.5%, the observed type I error rate should be 0.0250 ± 0.0031 with 95% probability. For simulating a true power of 50%, the observed power should be 0.5000 ± 0.0098 with 95% probability.

The simulations are summarized in Tables 1 and 2 and discussed in the following sections. Within the tables, success is the population probability of success, inv is the investigational device group, std is the control device group, N is the sample size at a site for the specified treatment, prob (reject) is the probability of rejecting the null hypothesis for the given method, and X indicates scenarios for which the population proportions are consistent with the null hypothesis for the given test.

3.1 Type I Error

Table 1 shows some results of type I error. Of note, the null hypothesis may be true for the FM hypothesis but false for the YTH hypothesis however, these situations are classified as being under the null in Table 1. As shown in Table 1, the type I error rate was maintained at or near the nominal level in most situations studied, with the observed type I error rate for the YTH method often being slightly larger than the observed type I error rate for the FM method, but generally within sample error of the target. The FM method was more often conservative, but there was not a qualitative difference between the two type I error rates with the exception of situations in which the success rates varied dramatically from site to

site. When success rates were 90% versus 80% at half of the sites and 50% versus 40% at the other sites (situation 8), the FM method was conservative while the YTH method was not. However, the degree of conservatism was small, about a percentage point or less. However, another situation is notable, one that results in important qualitative differences. When the success rates were 95% versus 85% at half of the sites and 45% versus 35% at half of the sites, and randomization was balanced with 25 subjects/treatment/site (situation 11), the FM test was slightly conservative (1.13%) while the YTH test was not (2.61%). However, with a minor imbalance in randomization (26 subjects versus 24 subjects in the two treatment arms at each site) the type I error rate for the FM test was affected dramatically: 0.26% to 3.16%, depending on the direction of the imbalance in each site (situation 12 & 13). Note that in each case, there were 150 subjects/treatment and 50 subjects/site, so the only difference was a minor imbalance within sites. The YTH test in these situations (12 & 13) controlled the type I error appropriately and consistently (2.55% to 2.62%). Finally, consider situations 9 and 10. In these situations, the null hypothesis is true under the FM framework but false under the YTH framework. However, the probability of rejecting the null hypothesis is not above the nominal 2.5% level for the YTH test after accounting for simulation variability.

Table 1: Simulated Type I Error Rates

Situation	Site	Success Std	Success Inv	N/site Std	N/site Inv	Prob (reject) FM	Prob (reject) YTH	Test of FM Null	Test of YTH Null
1	1-6	0.7	0.6	25	25	2.75%	2.86%	X	X
2	1-3	0.7	0.6	20	20	2.79%	2.85%	X	X
	4-6	0.7	0.6	30	30				
3	1-6	0.9	0.8	25	25	2.58%	2.51%	X	X
4	1-3	0.9	0.8	20	20	2.45%	2.43%	X	X
	4-6	0.9	0.8	30	30				
5	1-6	0.4	0.3	25	25	2.41%	2.49%	X	X
6	1-3	0.4	0.3	20	20	2.54%	2.65%	X	X
	4-6	0.4	0.3	30	30				
7	1-3	0.9	0.8	25	25	1.61%	2.73%	X	X
	4-6	0.5	0.4	25	25				
8	1-3	0.9	0.8	30	30	1.74%	2.67%	X	X
	4-6	0.5	0.4	20	20				
9	1-3	0.65	0.65	25	25	2.64%	2.72%	X	
	4-6	0.75	0.55	25	25				
10	1-3	0.45	0.85	25	25	1.12%	0.64%	X	
	4-6	0.95	0.35	25	25				
11	1-3	0.95	0.85	25	25	1.13%	2.61%	X	X
	4-6	0.45	0.35	25	25				
12	1-3	0.95	0.85	26	24	0.26%	2.62%	X	X
	4-6	0.45	0.35	24	26				
13	1-3	0.95	0.85	24	26	3.16%	2.55%	X	X
	4-6	0.45	0.35	26	24				

3.2 Power

As shown in Table 2, the power of the test was often similar between the two tests, generally with slight but unimportant differences favoring the YTH test. Exceptions were in situations in which the success rates varied strongly among sites. In these situations (21 & 22), the YTH test had higher power than the FM test. Also, as was observed with type I error rates, minor imbalances within studies had a major impact on the power of the FM test but negligible impact on the power of the YTH test (situations 23 25). In situations 27 through 29, a single site has success rates in favor of the investigational device while the remaining five sites have success rates in favor of the standard of care. Situation 28 was designed to have an overall success rate around 80% for the standard of care and 70% for the investigational device. This situation shows a large difference between the two tests and is a situation of concern. However, situation 29 has a similar set of parameter values, and the FM test has a slightly higher probability of rejecting the null hypothesis, indicating that the FM test does not completely solve the situation.

3.3 Simulation Conclusions

In conclusion, for data that exhibit similar success rates at the various sites and balanced randomization, either test performed acceptably well. When the data exhibit unequal success rates or very slight imbalances in randomization, the YTH test generally performed better, with better control of the type I error rate and higher power. Being more robust to unexpected but possible differences among sites, the YTH method is preferred. Finally, a single dramatically different site can confuse both methods.

4. Interaction Tests

Given the results observed when one site was dramatically different, we suggest that interaction tests may be desired. As noted in the simulation discussion, in scenarios where investigational is superior at half the sites and standard is superior at other half, both YTH and FM have high power to declare non-inferiority. These scenarios suggest that neither test adequately responds to qualitative treatment-by-site interaction.

Wiens and Heyse (2003) discuss interaction tests for non-inferiority trials. They discuss two types of interaction. Qualitative interaction is observing a point estimate consistent with non-inferiority in one stratum but not in another. Quantitative interaction is observing point estimates consistent with non-inferiority in all strata, but with different magnitudes of difference. Note that a conclusion of a quantitative interaction only may not invalidate conclusions from the combined analysis.

The interaction tests discussed can be calculated as

$$Q^- = \sum_{i=1}^{n} \left(D_i + \delta\right)^2 I\left(D_i > -\delta\right) / \sigma_i^2$$

$$Q^+ = \sum_{i=1}^{n} \left(D_i + \delta\right)^2 I\left(D_i < -\delta\right) / \sigma_i^2$$

$$Q = MIN\left(Q^+, Q^-\right)$$

where δ is the non-inferiority margin, D_i is the treatment group difference within subset i, σ_i^2 is the variance of the treatment difference within subset i, $I\left(D_i > -\delta\right)$ equals 1 if

Table 2: Simulated Power

Situation	Site	Success Std	Success Inv	N/site Std	N/site Inv	Prob (reject) FM	Prob (reject) YTH	Test of FM Null	Test of YTH Null
14	1-6	0.7	0.8	25	25	97.98%	98.01%		
15	1-3	0.7	0.8	20	20	98.06%	98.22%		
	4-6	0.7	0.8	30	30				
16	1-6	0.85	0.95	25	25	99.99%	99.99%		
17	1-6	0.45	0.55	25	25	94.04%	93.84%		
18	1-3	0.85	0.95	25	25	97.42%	99.43%		
	4-6	0.35	0.45	25	25				
19	1-3	0.35	0.95	25	25	97.46%	97.68%		
	4-6	0.85	0.45	25	25				
20	1-6	0.7	0.7	25	25	46.54%	46.94%		
21	1-3	0.5	0.5	25	25	47.96%	60.04%		
	4-6	0.9	0.9	25	25				
22	1-3	0.3	0.3	25	25	41.83%	64.87%		
	4-6	0.9	0.9	25	25				
23	1-3	0.3	0.3	24	26	22.85%	64.14%		
	4-6	0.9	0.9	26	24				
24	1-3	0.3	0.3	26	24	62.47%	65.91%		
	4-6	0.9	0.9	24	26				
25	1-3	0.3	0.3	30	30	42.06%	61.38%		
	4-6	0.9	0.9	20	20				
26	1-3	0.4	0.5	25	25	44.29%	44.62%		
	4-6	0.9	0.8	25	25				
27	1-5	0.65	0.55	25	25	41.94%	44.86%		
	6	0.45	0.95	25	25				
28	1-5	0.94	0.642	25	25	2.19%	37.69%	X	
	6	0.01	0.99	25	25				
29	1-5	0.8	0.7	25	25	95.24%	93.09%		
	6	0.01	0.99	25	25				

Table 3: YTH Example

Stratum	Success Investigational	Success Standard	Difference
1	13/23 (56.5%)	15/29 (51.7%)	4.8
2	30/50 (60.0%)	27/45 (60.0%)	0.0
3	19/38 (50.0%)	8/31 (25.8%)	24.2
Combined	62/111 (55.9%)	50/108 (47.6%)	8.3

$D_i > -\delta$ and 0 otherwise, and $I\left(D_i < -\delta\right)$ equals 1 if $D_i < -\delta$ and 0 otherwise. $Q > c$ or $Q^+ > c$ lead to conclusions of qualitative or quantitative interaction, respectively, with c from Gail and Simon (1985).

Using this interaction test on situation 19 in Table 2, we find significant quantitative and qualitative interactions. Both tests had high power in this situation to reject the null hypothesis; however, investigation of the differences within each site shows that there is a need for further exploration.

5. Example

Yanagawa, Tango and Hiejima demonstrated their method with a sample data set. This data set is reconstructed in Table 3.

As analyzed in the paper, the stratified analysis using the YTH method resulted in concluding non-inferiority with a margin of 5 percentage points (Z = 2.033; p = 0.0210). However, analysis with the unstratified FM method did not result in a conclusion of non-inferiority (Z = 1.949; p = 0.0257).

Confidence intervals were created corresponding to the YTH and the FM testing methods. The intervals contain all values of $\pi_D\pi_S$ that would not be rejected if used as the margin in the non-inferiority hypothesis (Sterne, 1954). The lower bound of a one-sided 97.5% confidence interval with the YTH method is 0.045 (or, 4.5%) indicating that any non-inferiority margin greater than or equal to 0.045 would result in a conclusion of non-inferiority. The corresponding lower bound with the FM method is 0.051.

The example illustrates that the stratified and unstratified analyses can provide different hypothesis testing conclusions, and that the resulting confidence intervals are informative. As mentioned previously, the confidence interval from the YTH method is calculated under the assumption that the difference in success rates is identical in each stratum, even if the underlying success rates are not. Thus, the confidence interval can be used to describe the common difference in success rates, taking stratification into account.

6. Summary and Discussion

The stratified analysis will be preferred if there is a substantial difference in success rates between sites. To guard against a situation in which not only the success rates but also the treatment effects differ dramatically among sites, an investigation of interaction should

be part of any analysis strategy. Situation 19 in Table 2, as an example, is a situation in which the investigational device is strongly superior to the standard of care in half the sites and strongly inferior in the other half. Both tests have high power to conclude non-inferiority, indicating that neither test adequately responds to the qualitative treatment-by-site interaction. Qualitative interaction is important, but cannot be detected reliably by either the YTH or FM method, and should be investigated separately from the primary assessment of non-inferiority.

REFERENCES

Farrington C.P. and Manning G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*. 9: 1447-1454.

Fisher R.A. (1956). The mathematics of the lady tasting tea. In Newman, J.R. *The world of mathematics, Volume III, Part VIII, Statistics and design of experiments*. Simon & Schuster: New York, New York, pp 1514-1521.

Gail, M., Simon, R. (1985). Testing for qualitative interaction between treatment effects and patient subsets. *Biometrics*. 41:361372.

Kahan B.C. and Morris T.P. (2011). Improper analysis of trials randomized using stratified blocks or minimisation. *Statistics in Medicine*. 31: 328-340.

Nicholas, K. (2014). "The Impact of Covariate Adjustment at Randomization and Analysis For Binary Outcomes: Understanding Differences Between Superiority and Non-Inferiority Trials." Presented at Society for Clinical Trials Annual Meeting, May 20, Philadelphia, PA.

Pocock S.J., Assman S., Enos L.E. and Kasten L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*. 21: 2917-2930.

Senn S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine*. 23:3729-3753.

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometricka*. 41: 275-278.

Wiens, B. L., and Heyse, J. F. (2003). Testing for Interaction in Studies of Noninferiority, *Journal of Biopharmaceutical Statistics*, 13:1, 103-115

Yanagawa T., Tango T. and Hiejima Y. (1994). Mantel-Haenszel-type tests for testing equivalence or more than equivalence in comparative clinical trials. *Biometrics*. 50: 859-864. (Correction in *Biometrics*. 51:392.)