Reducing Alpha Adjustment When Tests Are Structurally Correlated

Jonathan Siegel[1]
[1]Bayer HealthCare Pharmaceuticals Inc., 100 Bayer Boulevard, Whippany, NJ 07981

Abstract

Pharmaceutical clinical trials with multiple comparisons require adjustment for familywise false positive error. Positively correlated tests have less alpha inflation than independent tests. Dmitrienko, Tamhane, and Bretz (2010) outline general test procedures which can take advantage of alpha inflation reduction resulting from positive correlation. In many cases, the design itself results in structural correlation implicit in design assumptions. A customized adjustment taking into account the specific design context can result in greater power and/or reduced sample sizes. This presentation outlines the approach, briefly covers background theory, and discusses and provides theory-derived and/or simulation results for examples including a multi-arm trial with a common control arm, a trial with an overall population and a subgroup, and correlation between PFS and OS under a correlated bivariate exponential model.

Key Words: Multiple testing, Correlated test, Multi-arm trial, Shared control arm, subgroup, PFS and OS

1. Background

Pharmaceutical industry regulators require strong control of the family-wise error rate (FWER) when submitting trials with multiple endpoints and/or comparisons. The topic has attracted significant regulatory attention. In 2002, the EMEA issued “Points to Consider on Multiplicity Issues in Clinical Trials.” The FDA has been developing a draft “Guidance for Industry: Analysis of Clinical Trials with Multiple Endpoints”

Clinical trialists have sought to gain more from each trial by increasing the number of treatments and endpoints evaluated, and developing a more comprehensive picture of the effect of a treatment on patient survival, disease status, symptoms, and quality of life. These efforts have tended to increase the number of endpoints and hypotheses tested in each trial.

Because strong multiplicity control can be costly in study power, sample size, and duration, a variety of techniques have been developed to make multiplicity control more efficient. Efficient techniques have tended to focus on composite endpoints and on closed testing procedures with parallel, sequential, or hybrid testing techniques.

2. Correlated Tests

When tests are positively correlated, the family-wise error rate is less than under independence, as rejection of one null hypothesis increases the likelihood of rejection of others (and vice versa). Dmitrienko et al (2010) note that standard nonparametric testing procedures (e.g. Bonferroni, Holm) perform poorly in the presence of correlated data, Semi-parametric procedures (e.g. Hochberg, Hommel) perform better in the presence of strongly correlated data but are not optimal. Procedures based on specific distributional assumptions (e.g. Dunnett's Test) perform best.

Incorporating information about test correlation structure into the method for controlling the FWER can result in more efficient trial designs.

Several general techniques have been developed to evaluate and incorporate correlation structure during analysis, including the boostrapping approach of Westfall and Young (1993), and Romano and Wolf (2011) These techniques rely on assumptions which reduce the number of elements in the correlation matrix to a manageable number, including subset pivotality for Westfall and Young, and monotonicity of critical values for Romano and Wolf.

3. Structural Correlation

The approach proposed in this paper uses correlation structure inherent in a study design and its fundamental assumptions to reduce the degree of alpha adjustment, and hence the sample size, needed to achieve a given FWER at a given power. This approach uses study distributional assumptions and requires a method specific to a particular study design. It is not a general method and not appropriate for all trials.

Unlike general methods, it does not require introducing an array of new assumptions, and it is not dependent on the results of post-hoc analyses. It depends only on protocol and endpoint definition assumptions. The alpha adjustment can be calculated in advance and pre-specified in the protocol or SAP. Although the correlation characteristics and degree of alpha adjustment can in principal be derived from theory, simulations are required in practice. The needed simulations are often computationally intense.

Many common trial designs and endpoints have characteristics resulting in structural correlation between hypothesis tests, and hence potentially can be used to support a study-specific alpha adjustment. The use of structural correlation proposed here is not necessarily always acceptable for regulatory use. Even where its acceptability is questionable, it can still be used for evaluations of sponsor risk (study power and sample size), and might be usable in situations where regulatory leniency is possible (e.g. rare diseases, accelerated approval, etc.).

4. Three Examples

This paper presents three examples which should help illustrate the breadth of the potential applicability of the approach. These examples are:

- Correlation between tests on an overall study population, and on a subgroup.

- Correlation between the treatment vs. control hypothesis tests in a multi-arm study sharing a control arm.

- Correlation between PFS and OS due to the common death events.

5. Correlation Between Population and Subgroup

We'll use a single-arm response example for illustration

Suppose a population has two subgroups, Group 1 which is the subpopulation of interest and Group 2 comprising the remaining patients.

Let Group 1 have proportion q_1=q and response rate p_1, Y ~ Bin(n; q)
Let Group 2 have proportion q_2=1-q and response rate p_2

Then the total population can be modeled as a 2-component mixture of binomial populations

$$P(X = m) = \sum_{i=1}^{2} q_i \, p_i^m (1 - p_i)^{n-m}$$

The total population variance is accordingly

$$V_T = \left(\sum_{i=1}^{2} q_i [np_i(1 - p_i) + (np_i)^2]\right) - \left(\sum_{i=1}^{2} q_i np_i\right)^2$$

The correlation between population and subgroup response is thus

$$\frac{n[(1 - q_1p_1)(q_1p_1 + q_2p_2)]}{\sqrt{V_T}\sqrt{q_1p_1(1 - p_1)}}$$

In practice, particularly for multiple subgroups and/or use of a multi-stage design, the correlation between tests is calculated using simulations.

6. Multi-Arm Trials Sharing a Control

Multi-arm efficacy trials are used, especially in Phase II, to evaluate different formulations or regimens in a single trial. Sample sizes in Phase II are usually insufficient to support direct pairwise comparison of all treatments. Instead, hypothesis tests are performed only for comparisons with the control arm.

Wason and Jaki (2012) showed that under simplifying assumptions (e.g. the allocation to each experimental arm is equal), the correlation between normal test statistics at each stage is 1/(1 +r), where r is the proportion of the total sample size allocated to the control arm. Because of the complexity of the analytic forms as shown by Magirr et al. (2012), Wason and Jaki recommend simulations to establish trial operating characteristics.

Wason and Jaki propose a quicker method of simulating trials than generating each individual patient outcome. They propose generating matrices of test statistics for each comparison at each stage and using them to obtain simulated operating characteristics. They note a large number of replicates is necessary to reduce error in the estimates, and recommend 250,000 replicates for “a good estimate in practice”

In real-world practice, trial durations and interim stopping probabilities need to be estimated with non-constant accrual and drop-out patterns taken into account. Designs often have other complicating characteristics, and may require sensitivity analyses for contingencies such as cross-over or additional population heterogeneity. Accordingly, in evaluating a real large-scale trial, it will often be necessary to simulate at the individual-patient level to take these additional specification characteristics and sensitivity contingencies into account.

Simulating at the individual patient level with a sufficient number of replicates to obtain a reliable calculation can be computationally intensive. Multiple runs should be performed to evaluate computational replicability for the number of replicates chosen. Efficient programming and monitoring of computational resource use is critical. So long as the alpha is estimated accurately, other characteristics such as power, study duration, etc. can be and in practice will often need to be estimated with somewhat less than ideal reliability in order to reduce the number of replicates to manageable proportions. As underlying distributional, efficacy, and accrual assumptions are imperfect and often merely educated guesses, a small amount of imprecision (e.g. calculating power to nearest percent) is generally a reasonable tradeoff to obtain a smaller sample size.

7. Correlation Between PFS and OS

Progression-free survival (PFS) is a common time-to-event endpoint in Phase II oncology studies. PFS events are defined as occurring at the earlier of tumor progression or death. It is a composite of time to tumor progression (TTP) and overall survival (OS). Because death events are common to PFS and OS. the two endpoints are correlated. Taking the correlation into account requires making assumptions about their joint distribution.

Fleischer et al. (2009) proposed a model they characterized as the “maximal independence” model, making the minimum dependence assumptions necessary to share death events.

Under the model:

OS ~ Exp(λ_1)
TTP ~ Exp(λ_2)
PFS = min(OS, TTP)

It follows that PFS ~ Exp($\lambda_1+\lambda_2$)

Fleischer et al. showed that Corr(PFS, OS) = $\lambda_1/(\lambda_1+\lambda_2)$ = MedPFS/MedOS

In their paper Fleischer et al. discuss a more general model that takes into account additional assumptions about dependence between PFS and OS. For our purposes, however, we want to assume only structural correlation, dependence inherent in PFS and OS having death events in common.

The Fleischer maximal independence model represents a natural representation of structural correlation. It results in both PFS and OS having a simple exponential distribution, which is consistent with general protocol assumptions. It can be incorporated into patient-level clinical trial simulations in a straightforward way, with the FWER (probability of OS or PFS succeeding under joint H0) estimated without additional complexity.

8. Example Simulation

An example simulation study is presented, evaluating the correlation between a subgroup and the total study population for a survival analysis endpoint. The simulation scenarios shown use a subgroup proportion between 10% and 90%. Each simulated trial randomizes 500 simulated patients into subgroup vs. other and treatment vs. control groups. For the simulation performed under H0, all patients receive a simulated random survival time from the same exponential distribution.

The simulation performs a log-rank test for treatment vs. control separately for all patients, and for patients in the subgroup. A 2-sided log-rank p-value is obtained. The correlation coefficients between the subgroup and overall Kaplan-Meier medians and log-rank p-values are calculated. For each comparison, alpha is calculated as the proportion of log-rank trials for which the simulated p-value was < 0.05. The FWER is calculated as the proportion of trials at which at least one of the group and subgroup comparisons had a simulated p-value < 0.05. Simulations used SAS 9.2. 50,000 replicates were used. Results are shown in Table 1.

Table 1: Simulation Results

Percent in Subgroup	*Simulated Alpha Total Population*	*Simulated Alpha Subgroup*	*Simulated Correlation Between Kaplan-Meier Medians*	*Simulated Correlation Between Log-Rank p-values*	*Simulated FWER*	*Hochberg and Hommel FWER*
10%	0.0505	0.0568	0.282	0.055	0.1016	0.1049
20%	0.0518	0.0538	0.421	0.121	0.0974	0.1013
30%	0.0513	0.0530	0.521	0.206	0.0934	0.0981
40%	0.0524	0.0510	0.610	0.277	0.0889	0.0950
50%	0.0519	0.0523	0.683	0.363	0.0866	0.0925
60%	0.0515	0.0528	0.756	0.459	0.0835	0.0908
70%	0.0496	0.0505	0.820	0.560	0.0761	0.0849
80%	0.0511	0.0502	0.880	0.688	0.0721	0.0843
90%	0.0507	0.0509	0.939	0.823	0.0655	0.0802

9. Discussion

The simulated FWER was close to the nominal 0.10 for small subgroup proportions, but became noticeably smaller as the proportion of patients in the subgroup increases. Using the actual correlation structure resulted in a reduced total alpha compared to the Hochberg and Hommel procedures (which had identical results in this example). The Holm procedure (not shown) had nominal alpha of 0.10 (same as Bonferroni).

Consistent with Wason and Jaki's findings regarding multi-arm trials, a very large number of replicates is required for each simulation which may be beyond feasible computing capacity absent enhanced computing facilities. 50,000 replicates took over an hour for each overnight run on a Unix system. As indicated by the noticeable differences in simulated total population alpha across scenario runs, 50,000 replicates appeared inadequate for regulatory use, and Wason and Jaki's estimate of 250,000 replicates might be more appropriate.

To obtain a desired FWER value, simulations would need to be run iteratively, adjusting the nominal alpha values until the simulated FWER value reaches the desired value.

10. Conclusions

Structural correlation is potentially available for exploitation in many common clinical trial situations. Utilizing structural correlation can result in spending less alpha than standard nonparametric and semi-parametric procedures.

Simulation-based approaches based on fundamental, generally-accepted, pre-hoc protocol assumptions may be more reliable in a pharmaceutical-trial context than bootstrap methods based on post-hoc analyses of a single trial dataset.

However, the simulations needed to account for structural correlation are complex and extremely computationally intensive. The method is not appropriate for all situations. It may be infeasible absent enhanced computing resources.

The approach is not, however, without potential. Pivotal oncology trial costs per patient can run in the tens of thousands of dollars and potential sales can be in the hundreds of millions or billions. In this context, the sample size reduction and/or increased success chance of even a modest relaxation of the nominal FWER alpha requirements can justify the costs, especially as computing efficiency improves.

Acknowledgements

The author gratefully acknowledges Sylvia Engelen, Daniel Haverstock, and Minghua (Michael) Shan for encouragement and support in preparing this presentation and paper.

References

Dmitrienko, A, Tamhane, AC, and Bretz, F, Eds. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall (2010)

European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products. Points to Consider on Multiplicity Issues in Clinical Trials. EMEA 2002.

Fleischer, F, Gaschler-Markefski, B and Bluhmki, E. A statistical model for the dependence between progression-free survival and overall survival. *Statist. Med*. 28:2669–2686 (2009)

Magirr, T, Jaki, T and Whitehead, H, A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 99:494-501 (2012).

Romano, JP and Wolf, M. Exact and approximate stepdown methods for multiple hypothesis testing. *JASA* 100:94-108 (2005).

Wason, MS and Jaki, T. Optimal design of multi-arm multi-stage trials. *Statist. Med.* 31:4269-4279 (2012)

Westfall, PH and Young, SS. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley (1993).