A Versatile Multi-domain Test with Adaptive Weighting

Yang Zhao[1], Stephen Lake[2]
[1,2]Sanofi R&D, 640 Memorial Dr., Cambridge, MA 02139

Abstract

The design of a clinical trial to investigate a treatment for a rare disease is often complicated by the multi-systemic nature of the disease; no single endpoint can capture the spectrum of potential therapeutic benefits. Multi-domain outcomes which take into account patient heterogeneity of disease presentation through measurements of multiple symptom/functional domains are an attractive alternative to a single endpoint. To obtain the totality of evidence for treatment efficacy over endpoints from various disease progression domains, an extension of a test for equality of two survival distributions based on weighted differences of Kaplan-Meier curves [Uno *et al.* (2015)] is proposed. The test is a weighted sum of domain-specific test statistics with weights selected adaptively via a data-driven algorithm. The null distribution of the test is constructed empirically through resampling. We used data from clinical trials in a rare lysosomal storage disorder and in multiple sclerosis to illustrate the advantage of the combined testing procedure over the conventional methods. Simulations were conducted to demonstrate the statistical properties of the test and to compare to alternative methods.

Key Words: multivariate test, non-parametric test, rare disease, treatment effect, multi-domain outcome, adaptive weights

1. Introduction

In therapeutic areas such as rare disease and multiple sclerosis, the disease manifestation is often multi-systemic. For example, in MPS I, a mucopolysaccharide storage disorder, the deficiency of alpha-L-iduronidase can lead to the accumulation of glycosaminoglycans in a wide variety of tissues, thus a broad spectrum of clinical symptoms including cardiac disease, respiratory disease, joint stiffness, developmental delay, etc. In multiple sclerosis, a demyelinating disease of the central nervous system, the focal tissue injury of the brain and spinal cord can result in a constellation of chronic clinical symptoms including muscle weakness, impaired mobility, bladder/bowel dysfunction, cognitive and visual impairments, etc. Therefore, to investigate a treatment in these areas, no single endpoint can capture the spectrum of potential therapeutic benefits. Instead, multi-domain outcomes are an attractive alternative to obtain the totality of evidence for treatment efficacy over endpoints from various disease progression domains.

There has been a rich literature on multi-domain tests. O'Brien (1984) proposed a generalized least squares (GLS) test and a nonparametric rank-sum test that extends the Wilcoxon rank-sum test in the multi-domain case. Wei and Lachin (1984) described a class of multivariate asymptotically distribution-free tests for incomplete multi-domain observations. The Wei-Lachin test applies to censored time-to-event data as well as

missing completely at random ordinal data. A comprehensive overview of estimators and tests for multivariate partially incomplete data from two populations is given by Lachin (1992), in which the Wei-Lachin test, Wei-Johnson test, GLS test, etc. were discussed and compared. More recently, Xu *et al.* (2003) proposed a test with adaptive weighting that combines dependent tests for linkage across multiple phenotypic traits. Asymptotic normality of the dependent test statistics is required and their covariance matrix can be estimated in the context of a linkage study in genetic epidemiology. Uno *et al.* (2015) developed a versatile test with similar adaptive weighting for equality of two survival functions based on weighted differences of Kaplan-Meier curves. A perturbation resampling method was utilized to empirically approximate the limiting distribution of the test statistics. In this paper, we extend the test with adaptive weighting for the multi-domain outcome setting in the clinical trials context, and propose to use a permutation-based procedure for statistical inferences. Our objective is to evaluate the utility of this test in analyzing clinical trial outcomes, and to compare it with conventional multi-domain testing methods.

This paper is organized as follows. In Section 2, we describe the specifics of the permutation test with adaptive weighting. A simulation study is presented in Section 3, and real data analysis of two clinical trials is discussed in Section 4. We summarize our findings and conclude in Section 5.

2. The Permutation Test with Adaptive Weighting

2.1 The Framework

Assume that in a clinical trial, two treatment groups are to be compared based on K domains of continuous outcomes. Let $\boldsymbol{Y}_{ij} = (Y_{i1j}, \cdots, Y_{iKj})$ denote the outcomes from the K domains for the j^{th} patient in the i^{th} treatment group, where $i = 0,1$ represents the control or the treatment group, and $j = 1, \cdots, n_i$. Assume $\boldsymbol{Y}_{ij}$'s are independent random vectors with distribution functions $F_i(y_1, \cdots, y_K)$. Without loss of generality, assume that for all K domains, a larger value of an outcome indicates a better clinical benefit.

The statistical hypothesis of interest is the following:

$$H_0: F_1(y_1, \cdots, y_K) = F_0(y_1, \cdots, y_K)$$
$$H_1: F_1(y_1, \cdots, y_K) \succ^s F_0(y_1, \cdots, y_K)$$

for all $(y_1, \cdots, y_K) \in \mathbb{R}^K$. Here we wish to detect the stochastic ordering of two multivariate distributions, that is, $F_{1k}(\cdot) \leq F_{0k}(\cdot)$ for each marginal distribution function F_{ik} of F_i, $i = 0,1$, $k = 1, \cdots, K$, with at least one strict inequality.

2.2 Testing Procedure

We extend the testing methods of Xu *et al*. (2003) and Uno *et al*. (2015) to a multi-domain test setting in a general clinical trial context and propose to use a permutation-based procedure for statistical inferences. As a key assumption in Xu *et al*. (2003) and Uno *et al*. (2015), asymptotic joint normality (process) of the marginal statistics (process) requires that the covariance be estimated consistently and accurately. It was possible to estimate the covariance matrix in the genetic epidemiology linkage studies, or to empirically approximate the limiting distribution using a perturbation resampling technique for the Kaplan-Meier survival process. However, in the general clinical trial setting, it may not always be straightforward to estimate the covariance, or to estimate it accurately. Therefore, we use a permutation procedure to bypass this difficulty and preserve the dependence structure.

Let Z_k be the marginal test statistic comparing the treatment group versus control for the outcomes of the k^{th} domain, $k = 1, \cdots, K$. Define $V(c) = \sum_{k=1}^{K} W_k(c) Z_k$, where $W_k(c) = \max\{Z_k, c\}$, $c \in [0, \eta]$ is a data-driven parameter, and η is a pre-specified parameter. The test procedure below constructs a test statistic based on $\{V(c), 0 \le c \le \eta\}$ and chooses c adaptively (similar to Xu et al. 2003, Uno et al. 2015):

1. Simulate the null joint distribution of $\boldsymbol{Z} = (Z_1, \cdots, Z_K)$ via permutation, hence the approximated null distribution of $V(c)$ indexed by c (reference set $\boldsymbol{D}$);
2. Let $v(c)$ be the observed value of $V(c)$, its p-value $p(c)$ can be obtained for each c;
3. Let p_b be the most significant $p(c)$, i.e. $p_b = \min\{p(c): c \in [0, \eta]\}$;
4. Let $P(c)$ and P_b be the random counterpart of $p(c)$ and p_b, and take P_b as the test statistic;
5. The null distribution of P_b can be approximated by generating a large number of $\boldsymbol{Z}$'s via permutation: for each realized $\boldsymbol{Z}$, compute $V(c)$ and use reference set $\boldsymbol{D}$ to obtain the corresponding $P(c)$ and P_b;
6. The p value based on test statistic P_b is given by $pr(P_b < p_b)$.

The parameter η in the above procedure can be any pre-specified positive constant. Extensive simulation studies show that $\eta = 4$ should be a reasonable choice, which provides stable results, as the Z_k's would rarely be larger than 4 under H_0.

Heuristically, a good test should possess the property that under H_0, the distribution of the test statistic has a relatively short tail, but under a general one-sided alternative hypothesis H_1, a long tail, so that the observed statistic is likely to be large and thus reject H_0. Under the normal distribution assumptions, a test statistic constructed as a linear combination of the Z_k's has a short tail under H_0, but has low power against a general one-sided H_1; a test statistic using Z_k itself as the weight has a fat tail chi-squared distribution under H_0, hence not very powerful against specific alternatives. The proposed weighted sum $V(c) = \sum_{k=1}^{K} \max\{Z_k, c\} Z_k$ is a flexible statistic in that under H_0, it behaves like a linear combination of the Z_k's with a short tail, while under H_1, behaves like a long tailed chi-squared statistic.

2.2.1 Marginal test statistic Z_k

Since the adaptive weight is defined as $W_k(c) = \max\{Z_k, c\}$, $c \in [0, \eta]$, Z_k and c should be of comparable magnitude to ensure sensitivity. Although our test procedure does not require a joint normal distribution of the Z_k's, it is advisable to use normal-like statistics to achieve the aforementioned desired property. Convenient choices include the student t statistic or the standardized Wilcoxon rank-sum statistic. Both will be explored in the simulation study (Section 3) and the real data analysis (Section 4).

2.2.2 Choice of the threshold parameter c

For a fixed c, say $c = 2$, under H_0, since Z_k is approx. standard normal, $W_k(c) \approx 2$ for most k's, and so $V(c)$ should have a short tail; under H_1, a large Z_k results in $W_k(c) \approx Z_k$ and the observed $V(c)$ would be large. It is not clear though that $c = 2$ would still be a good choice when most of the Z_k's are positive but not large. If we set $c = 0$, then $V(c)$ has a long tail and behaves like a chi-squared statistic; or if $c = 4$, $V(c)$ has a short tail and behaves like a linear combination statistic. Therefore, it is not straightforward as to how to choose a fixed value of c a priori. The test procedure above provides an automatic

and objective way to adaptively choose parameter c, so that better power may be achieved with the flexibility.

3 Simulation Study

3.1 Simulation Setting

In this section, we demonstrate our proposed method with the simulated data. In the simulation, we explore two endpoints in a randomized placebo-controlled clinical trial. The clinical outcomes were generated from bivariate normal distribution with various treatment effect size assumptions. Specifically, we simulate two continuous, normally distributed outcomes with variance 1 and correlation ρ =-0.8, -0.4, -0.2, 0, 0.2, 0.4, 0.8. Treatment effect for each outcome takes the values of 0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6. So there are 45 unique treatment effect combinations for the two outcomes. We consider $n = 25$ or $n = 50$ per arm as hypothetical trials for rare disease. Set $\eta = 4$ and c is adaptively determined along an equally spaced grid from 0 to η with 50 possible values. Simulations were repeated 500 times for each simulation scenario. For higher accuracy, type I error rate was estimated from 1000 runs. We compare O'Brien test (OB), Wei-Lachin test with Gehan weights (WLWX), with log-rank weights (WLLR), permutation test with adaptive weighting using student t statistic (SNT) and using Wilcoxon rank-sum statistic (SWX).

3.2 Simulation Results

The type I error rates are reported in Table 1. WLLR test tends to inflate the type I error rate the most, although the inflation shrinks as the sample size increases. OB test controls the type I error reasonably well with a few exceptions, and its performance does not seem to improve with a larger sample size. When the sample size is 25 per arm, permutation tests SWX and SNT have type I error slightly exceed the nominal level in several cases, but they preserve the error rate well for moderate sample sizes when n=50 per arm.

Table 1: Type I Error Rate

n=25 per arm					
$\alpha = 0.05$	$\rho = -0.8$	$\rho = -0.4$	$\rho = 0$	$\rho = 0.4$	$\rho = 0.8$
OB	0.040	0.046	0.041	0.052	0.046
WLWX	0.046	0.057	0.049	0.059	0.058
WLLR	0.060	0.063	0.064	0.062	0.064
SWX	0.051	0.047	0.051	0.054	0.053
SNT	0.046	0.042	0.049	0.053	0.053
n=50 per arm					
$\alpha = 0.05$	$\rho = -0.8$	$\rho = -0.4$	$\rho = 0$	$\rho = 0.4$	$\rho = 0.8$
OB	0.050	0.048	0.053	0.050	0.055
WLWX	0.047	0.048	0.058	0.051	0.053
WLLR	0.054	0.048	0.048	0.053	0.060
SWX	0.051	0.045	0.046	0.048	0.050
SNT	0.045	0.046	0.049	0.046	0.051

Power of the tests when n=25 per arm is presented in Figure 1. Here the treatment effect values 0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6 corresponds to scenario codes 1, 2, 3, 4, 5, 6, 7, 8, 9. For example, scenario 15 at the lower right corner in Figure 1 represents the scenario when the treatment effect is set as 0 for one domain and 0.8 for the other domain. Only scenarios up to scenario 55 are depicted for illustration.

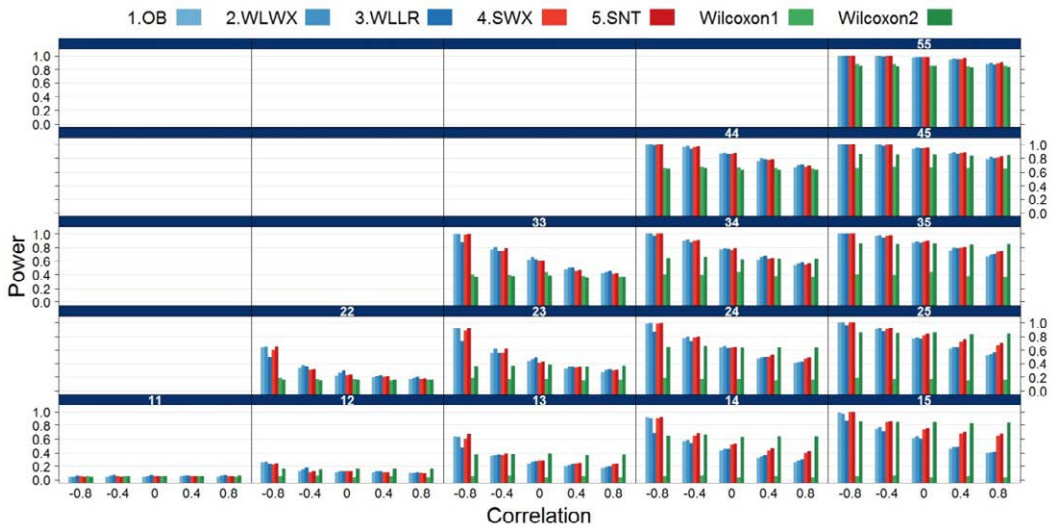

Figure 1: Power of Multi-domain Tests and Marginal Wilcoxon Rank-sum Tests (n=25 per arm)

When we have little prior knowledge about the treatment effect in each domain, it is obvious that using marginal Wilcoxon rank-sum test may not be ideal as the power can be extremely low for a domain with small treatment effect. WLLR shows clear inferiority compared to WLWX and OB tests in most scenarios. Permutation tests SNT and SWX perform as well as WLWX test in almost all scenarios, and better in cases when the treatment effect of one domain is relatively small while that of the other domain is large. Table 2 reports the power of WLWX and SWX when $\rho = 0.8$ for illustration. The performance pattern is similar for different correlation values, though the power across all tests tends to be higher.

Table 2: Power of Multi-domain Tests and Marginal Wilcoxon Rank-sum Test (n=25 per arm, $\rho = 0.8$).

Treatment effect	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
Wilcoxon test	0.058	0.16	0.37	0.63	0.84	0.96	0.99	1	1
0	0.058/ 0.053	0.11/ 0.11	0.19/ 0.23	0.28/ 0.40	0.41/ 0.64	0.53/ 0.85	0.64/ 0.96	0.76/ 0.99	0.85/ 1.00
0.2		0.19/ 0.17	0.31/ 0.30	0.42/ 0.47	0.54/ 0.67	0.67/ 0.85	0.79/ 0.95	0.88/ 0.99	0.93/ 1.00
0.4			0.43/ 0.41	0.56/ 0.54	0.69/ 0.73	0.82/ 0.88	0.89/ 0.96	0.94/ 0.99	0.96/ 1.00
0.6				0.69/ 0.67	0.81/ 0.81	0.89/ 0.91	0.95/ 0.97	0.97/ 1.00	0.99/ 1.00
0.8					0.90/ 0.88	0.95/ 0.95	0.98/ 0.98	0.99/ 1.00	1.00/ 1.00
1.0						0.98/ 0.98	0.99/ 0.99	1.00/ 1.00	1.00/ 1.00
1.2							1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
1.4								1.00/ 1.00	1.00/ 1.00
1.6									1.00/ 1.00

4 Application To Clinical Trials

In this section, we apply the multi-domain tests to two clinical trial data sets. Both are Sanofi Genzyme sponsored phase 3 studies.

4.1 MPS I Trial

This is a phase 3, randomized, double-blind, placebo-controlled clinical study of recombinant human alpha-L-iduronidase (rhIDU) in patients with mucopolysaccharidosis I (MPS I). This trial includes patients 5 years of age or older who were capable of standing independently for 6 minutes, and walking at least 5 meters within 6 minutes. Weekly intravenous infusions were administered for 26 weeks. The efficacy endpoints of interest are the changes from baseline at week 26 of the forced vital capacity, 6-minute walk test, AHI (apnea/hypopnea during sleep) index, and shoulder flexion. In total 45 patients were enrolled and treated, of which 37 patients (18 on treatment, 19 on placebo) had complete data on the 4 domains. We base our analysis on these 37 patients.

The results are summarized in Table 3. If the four hypotheses were tested separately, only the first endpoint (forced vital capacity) had a significant p value. Multiple testing procedures such as the fixed sequence test and the Bonferroni-Holm procedure would reject the first hypothesis while controlling for the family-wise-error-rate (FWER) of 0.025 for the one-sided hypothesis. Using the multi-domain testing methods, the overall hypothesis of equal joint distributions is rejected consistently. OB test seems less powerful than others. WLLR produced the smallest p value, but from the simulation study, it also tends to inflate the type I error rate. Permutation tests yielded reasonably small p values.

Table 3: Testing Results for MPS I Trial (N=37)

Marginal Test	**Wilcoxon p value**	**t test p value**
H_1: Forced vital capacity	0.0036	0.0021
H_2: 6-minute walk	0.0766	0.0772
H_3: AHI index	0.1850	0.1160
H_4: Shoulder flexion	0.4098	0.3778
Multiple Testing	**Rejected H_0**	**Rejected H_0**
Fixed sequence test	H_1 only	H_1 only
Bonferroni-Holm procedure	H_1 only	H_1 only
Multi-domain Test	**p value**	
O'Brien (OB)	0.0109	
Wei-Lachin with Gehan weight (WLWX)	0.0030	
Wei-Lachin with Log-rank weight (WLLR)	0.0008	
Adaptive weighting test using marginal student t statistic (SNT)	0.0067	
Adaptive weighting test using marginal Wilcoxon statistic (SWX)	0.0092	

4.2 Multiple Sclerosis

This trial is a phase 3, randomized, rater-blinded study comparing two annual cycles of intravenous alemtuzumab to three-times weekly subcutaneous interferon beta-1a (Rebif) in treatment-naïve patients with relapsing-remitting multiple sclerosis. The traditional clinical disability outcome for multiple sclerosis trials is the Expanded Disability Status Scale (EDSS) score. However, the EDSS score has now known to be insensitive to certain disability deteriorations, and is subject to low reliability between/within raters. In light of the limitations of the EDSS scores, additional measurements such as the timed 25-foot walk test and the 9-hole Peg test have been incorporated to enhance the EDSS measurement. Therefore, we consider four endpoints of interest: the changes from baseline at month 24 of the EDSS score, the timed 25-foot walk test, and the 9-hole Peg test using dominant or non-dominant hand. 581 patients were treated with alemtuzumab 12mg/day or Rebif, of which 533 (361 alemtuzumab and 172 Rebif) had complete data on the four endpoints.

Table 4 shows the testing results. Marginally, the EDSS failed to show statistical significance comparing alemtuzumab to Rebif, but the timed 25-foot walk and the 9-hole peg test with non-dominant hand showed significance. Nevertheless, controlling for the FWER at a nominal level 0.025, none of the four hypotheses can be rejected by either the fixed sequence test or the Bonferroni-Holm procedure. The advantage of using multi-domain tests is clear here, as WLLR and SWX yielded p values smaller than 0.025, thus statistically significant results. From the simulation study, WLLR tends to inflate the type I error rate, and the permutation tests SNT and SWX should be the preferred tests. In this particular example, SNT gives a non-significant result, which might be due to the fact that some of the endpoints (e.g. 9-hole Peg test) are highly skewed in distribution.

Table 4: Testing Results for Multiple Sclerosis Trial (N=533)

Marginal Test	**Wilcoxon p value**	**t test p value**
H_1: EDSS	0.8598	0.6719
H_2: Timed 25-foot walk	0.0179	0.2586
H_3: 9-hole Peg test dominant hand	0.0983	0.0528
H_4: 9-hole Peg test non-dominant hand	0.0134	0.0081
Multiple Testing	**Rejected H_0**	**Rejected H_0**
Fixed sequence test	None	None
Bonferroni-Holm procedure	None	None
Multi-domain Test	**p value**	
O'Brien (OB)	0.0855	
Wei-Lachin with Gehan weight (WLWX)	0.0433	
Wei-Lachin with Log-rank weight (WLLR)	0.0233	
Adaptive weighting test using marginal student t statistic (SNT)	0.0684	
Adaptive weighting test using marginal Wilcoxon statistic (SWX)	0.0210	

5 Conclusion

In this paper, we propose a permutation test for multi-domain outcomes in clinical trials based on Xu *et al*. (2003) and Uno *et al*. (2015). Our method does not require estimation of the covariance among the marginal test statistics, and therefore can be applied freely in a very general multi-domain outcome setting. The simulation study demonstrated that the proposed test performs as well as conventional multi-domain tests in almost all scenarios, and better in cases when the treatment effect of one domain is relatively small while that of the other domain is large. The type I error rate is also well controlled. We analyzed data from two clinical trials in MPS I (a rare disease) and in multiple sclerosis. Our permutation test with adaptive weighting was powerful in detecting treatment benefits, especially in the multiple sclerosis case where the multiple testing procedures failed to show statistical significance due to the large number of endpoints and moderate marginal *p* values.

Our simulation study focused on normal distributions. Future research may explore the performance of the tests under skewed distributions or with outliers. More than two endpoints may also be considered in a simulation setting to learn about the statistical properties of multi-domain tests under various treatment effect configurations.

Acknowledgements

The authors would like to thank professor L. J. Wei for insightful discussions and guidance.

References

1. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; 28: 586-604.
2. Dallow N.S., Leonov S. L., Roger J. H. Practical usage of O'Brien's OLS and GLS statistics in clinical trials. *Pharmaceutical Statistics* 2008; 7: 53-68. Doi: 10.1002/pst.268.
3. Lachin J.M. Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations. *Statistics in Medicine* 1992; 11:1151-1170.
4. Lachin J.M. Application of the Wei-Lachin multivariate one-sided test for multiple outcomes on possibly different scales. *PLoS ONE* 2014; 9(10):e108784. Doi:10.1371/journal.pone.0108784.
5. O'Brien P. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; 40:1079-1087.
6. Uno H, Tian L, Claggett B, Wei L.J. A versatile test for equality of two survival functions based on weighted differences of Kaplan-Meier curves. *Statistics in Medicine* 2015; DOI:10.1002/sim.6591.
7. Wassmer G, Reitmeir P, Kieser M, Lehmacher W. Procedures for testing multiple endpoints in clinical trials: an overview. *Journal of Statistical Planning and Inference* 1999; 82: 69-81.
8. Wei L.J. and Lachin J.M. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* 1984; 79(387):653-661.
9. Xu X, Tian L, Wei L.J. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* 2003; 4(2):223-229.