frontiers in
PUBLIC HEALTH

TECHNOLOGY REPORT
published: 09 December 2013
doi: 10.3389/fpubh.2013.00063

Multiple hypotheses testing procedures in clinical trials and genomic studies

Qing Pan*

Department of Statistics, The George Washington University, Washington, DC, USA

Edited by:
Zhiwei Zhang, Food and Drug Administration, USA

****Correspondence:***
Qing Pan, Department of Statistics, The George Washington University, 801 22nd Street NW, Rome Hall 665, Washington, DC 20052, USA
e-mail: qpan@gwu.edu

We review and compare multiple hypothesis testing procedures used in clinical trials and those in genomic studies. Clinical trials often employ global tests, which draw an overall conclusion for all the hypotheses, such as SUM test, Two-Step test, Approximate Likelihood Ratio test (ALRT), Intersection-Union Test (IUT), and MAX test. The SUM and Two-Step tests are most powerful under homogeneous treatment effects, while the ALRT and MAX test are robust in cases with non-homogeneous treatment effects. Furthermore, the ALRT is robust to unequal sample sizes in testing different hypotheses. In genomic studies, stepwise procedures are used to draw marker-specific conclusions and control family wise error rate (FWER) or false discovery rate (FDR). FDR refers to the percent of false positives among all significant results and is preferred over FWER in screening high-dimensional genomic markers due to its interpretability. In cases where correlations between test statistics cannot be ignored, Westfall-Young resampling method generates the joint distribution of *P*-values under the null and maintains their correlation structure. Finally, the GWAS data from a clinical trial searching for SNPs associated with nephropathy among Type 1 diabetic patients are used to illustrate various procedures.

Keywords: false discovery rate, family wise error rate, global test, multiple hypotheses testing, resampling method, stepwise procedure

1. INTRODUCTION

When more than one hypotheses are tested at the same time, it is well known that the family wise type I error rate (FWER), that is, the probability of reporting at least one significant finding when the null hypotheses are true, will be inflated. Take J independent test statistics as an example. When each test controls its type I error rate at α level, the FWER is $1-(1-\alpha)^J$. **Table 1** lists the FWERs for different combinations of J and α. When $J=10$ and $\alpha=0.05$, FWER goes up to 0.401. In cases of 100 or more simultaneous tests, it is almost sure to get false positive results.

Multiple hypotheses testing arises frequently both in clinical trials and in genomic studies. The different goals in these two settings result in different strategies. First, the hypotheses in clinical trials are often considered as a whole while those in genomic studies are more independent from each other. In clinic trials, multiple hypotheses are often considered jointly with a coherent theme. A few examples are given as follows. The symptoms of a complex disease often show up in different parts of the body or in different forms, such as different types of cancer. Multiple laboratory measurements monitor the underlying disease process, such as the viral loads and CD4 cell counts in HIV positive subjects. A treatment might have different responses in different patient subpopulations. On the other hand, multiple hypotheses in genomic studies arise because a large number of candidate markers are tested at the same time. Based on the number of tests carried out in the procedure, the multiple testing adjustment approaches can be grouped into global tests and stepwise procedures (1). Global tests summarize information from all endpoints/measurements/strata in one test statistic, while stepwise procedures carry out one test for each hypothesis. Therefore global tests are employed frequently in clinical trials while genomic studies almost always employ stepwise procedures. Second, the hypotheses in clinical trials are usually more specific with abundant prior information. In testing a specific treatment, with knowledge on the direction of the effects, directional tests with higher power can be employed. On the other hand, the genomic, epigenomic, transcriptomic, and proteomic network is much more complicated and often researchers screen for any signal, without knowing its direction or relationship to other markers. Third, the numbers of hypothesis in clinical trials are on a much smaller scale compared to the numbers in genomic studies – the numbers in clinical trials are usually less than ten, while the numbers of potential markers in genomic studies are sometimes over a million. In this manuscript, common procedures of multiple hypotheses adjustment in the two different settings are reviewed and compared.

The effects of interests are usually inferred from regression coefficients. In linear regression for normally distributed outcomes, the coefficient represents the difference in the outcome values between the groups being compared. In generalized linear models with logit link for binary outcomes, the coefficient equals the logarithm of the odds ratio of the outcome in the treatment group relative to the control group. In Cox proportional hazards models for partially censored failure time data, the exponentiated coefficient represents the hazards ratio. This review focuses on the choice of proper multiple testing adjustment method after the estimation procedures. Hence, we assume that appropriate models are chosen for different data configurations and parameters and covariance matrix are consistently estimated. Suppose there are J hypotheses in total.

Table 1 | FWER versus number of tests and the size of individual tests.

α	J	FWER
0.01	2	0.020
0.01	5	0.049
0.01	10	0.096
0.01	100	0.634
0.01	1000	1.000
0.05	2	0.098
0.05	5	0.226
0.05	10	0.401
0.05	100	0.994
0.05	1000	1.000

Let β and $\widehat{\beta}$ denote the two $J \times 1$ vectors of regression coefficients and their estimates, respectively, one element for each hypothesis. Furthermore, $\beta_j = 0$ corresponds to the jth null hypothesis, $j = 1, \ldots, J$.

2. MULTIPLE TESTING PROCEDURES IN CLINICAL TRIALS

2.1. SUM TEST

O'Brien (2) proposed a test derived from the generalized least squares principle

$$\sqrt{n}\mathbf{J}'\Sigma^{-1}\widehat{\beta},$$

where $\mathbf{J}$ is an $J \times 1$ vector of 1's and Σ is the covariance matrix of $\widehat{\beta}$. When elements of $\widehat{\beta}$ are independent from each other, the O'Brien test statistic reduces to a linear combination of $\widehat{\beta}_j$ where each $\widehat{\beta}_j$ is weighted by inverse of its variances. Tests employing linear combinations of $\widehat{\beta}_j$ with different weights have been proposed (3–6), among which the SUM test is especially popular (7). The SUM test statistic has a simple sum form

$$SUM = \sum_{j=1}^{J} \widehat{\beta}_j.$$

Under the null hypothesis $\beta_1 = \cdots = \beta_j = 0$, $E(SUM) = 0$. The SUM test is found to maximize the minimum power (maxmin test) for alternatives where all elements of β have the same sign (8, 9).

2.2. TWO-STEP

When homogeneous effects are of interests, a two-step procedure is commonly used. In the first step, we test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_j$ versus H_a: at least one pair $\beta_j \neq \beta_j'$ for $j \neq j'$ through Breslow-Day test or likelihood ratio test (LRT) (10, 11). Under the null, the LRT test statistic follows a Chi-square distribution with $J-1$ degree of freedom asymptotically. If the null hypothesis of homogeneous treatment effects is not rejected, we proceed to the second step where data from different endpoints are pooled and an overall treatment effect is estimated and tested against zero with a Wald test. The second test is carried out conditionally on the acceptance of the null in the first step. When the type I error rates in the two steps, α_1 and α_2, both equal 0.05, the marginal probability that the Two-Step procedure concludes homogeneous non-zero treatment effects under H_0 is $95\% \times 5\% = 4.75\%$, while the probability of concluding non-zero treatment effect in at least one endpoint under H_0 is $95\% \times 5\% + 5\% = 9.75\%$. Lachin and Wei (12) proposed to adjust α_1 and α_2 so that the overall type I error rate is $\alpha_1 + \alpha_2(1-\alpha_1) = 0.05$.

2.3. APPROXIMATE LIKELIHOOD RATIO TEST

The Hotelling's T test examines whether the vector β is a vector of zero

$$n\widehat{\beta}\Sigma^{-1}\widehat{\beta},$$

Here n is the sample size in testing each hypothesis. Under H_0, the Hotelling's T test statistic has an asymptotic Chi-square distribution. Follmann (13) modified Hotelling's T test for one-sided alternatives. His procedure rejects the null when the p-value of the Hotelling's T test is less than twice its nominal level and the sum of the treatment effects is in the desired direction (positive or negative). Tang et al. (14) proposed an approximate likelihood ratio test (ALRT) for one-sided alternative hypotheses. A $J \times J$ matrix A which satisfies $A'A = \Sigma^{-1}$ and $A\Sigma A' = I$ is calculated, where I denotes the identity matrix. Define $z = \sqrt{n}A\widehat{\beta}$ where the vector $\widehat{\beta}$ is mapped into a new vector z with independent components z_j, $j = 1, \ldots, J$. For H_a: at least one $\beta_j > 0$, the ALRT statistic is calculated as

$$ALRT = \sum_{j=1}^{J} \max(z_j, 0)^2,$$

where negative z_j values contribute zero. Hence the absolute magnitude of negative z_j has no impact on ALRT. The ALRT statistic follows a mixed Chi-square distribution under H_0.

2.4. MAX TEST

Another type of global tests employ the maximum of the standardized test statistics (15). The test statistic goes as follows

$$MAX = \max\left(\frac{|\widehat{\beta}_1|}{SD(\widehat{\beta}_1)}, \frac{|\widehat{\beta}_2|}{SD(\widehat{\beta}_2)}, \ldots, \frac{|\widehat{\beta}_J|}{SD(\widehat{\beta}_J)}\right),$$

where $SD(\widehat{\beta}_j)$ is the standard deviation of $\widehat{\beta}_j$. Given the one-to-one relationship between $\frac{|\widehat{\beta}_j|}{SD(\widehat{\beta}_j)}$ and its p-value, an equivalent test statistic is the minimum of the P-values. The MAX test is powerful to detect alternatives where the treatment effects are non-zero in at least one endpoint/measurement/stratum.

2.5. INTERSECTION-UNION TEST

Establishment of bioequivalency is required by the U.S. Food and Drug Administration (FDA) in approving generic drugs. The brand-name drug and its generic version are considered indifferent for the jth outcome if $\beta_j \in (-\varepsilon_j, \varepsilon_j)$, where the indifferent range ε_j is decided clinically. FDA is interested in whether the generic drug is superior in at least one aspect while non-inferior in all aspects. Therefore, the alternative of interest goes as follows H_a: $\{\max(\beta_1, \beta_2, \ldots, \beta_k) > 0\} \cap \{\min(\beta_1 + \varepsilon_1, \beta_2 + \varepsilon_2, \ldots, \beta_k + \varepsilon_k) > 0\}$

Table 2 | Comparison of five global test statistics.

Test	Test statistic
SUM	$SUM = \sum_{j=1}^{J} \widehat{\beta}_j$
Two-step: step one	$LRT = -2(L_0 - L_a)$
Two-step: step two	$\widehat{\beta}$
ALRT	$ALRT = \sum_{j=1}^{J} \max(Z_j, 0)^2$
IUT	$ALRT,\ T_j = \frac{\widehat{\beta}_j + \varepsilon_j}{SD(\widehat{\beta}_j)}\ j = 1, \ldots, J$
MAX	$MAX = \max\left(\frac{\widehat{\beta}_1}{SD(\widehat{\beta}_1)}, \frac{\widehat{\beta}_2}{SD(\widehat{\beta}_2)}, \ldots, \frac{\widehat{\beta}_J}{SD(\widehat{\beta}_J)}\right)$

L_0 and L_a represent the maximum log likelihood under H_0 and H_a respectively. And $\widehat{\beta}$ is the coefficient estimates from the pooled data.

where ∩ denotes intersection. The intersection-union test (IUT) (16–18) is most frequently used in these settings. It is a closed procedure which rejects the overall null hypothesis if and only if all null hypotheses included in the procedure are rejected. The *ALRT* is used to test against the alternative $max\ (\beta_1, \beta_2, \ldots, \beta_k) > 0$. Non-inferiority in the jth endpoint is tested by

$$\frac{\widehat{\beta}_j + \varepsilon_j}{SD(\widehat{\beta}_j)} \text{ for } j = 1, \ldots, J.$$

Because the overall rejection region is the intersection of all rejection regions, the overall type I error will not exceed α if the type I error rates of individual tests are set at α. Although more than one tests are carried out in IUT, it is included in the category of global tests because it draws an overall conclusion, not multiple hypothesis-specific conclusions. The five global tests are summarized in **Table 2**.

2.6. COMPARISON OF REJECTION REGIONS OF THE GLOBAL TESTS

We take the special example with two coefficient estimates $(\widehat{\beta}_1, \widehat{\beta}_2)$, which are are bivariate normal with mean (0, 0), variance 1 and 2 respectively, and correlation coefficient 0.3. The null and alternative hypotheses are H_0: $(\beta_1 = 0) \cap (\beta_2 = 0)$ versus H_a: $(\beta_1 > 0) \cap (\beta_2 > 0)$. The rejection regions of the five global tests are shown in **Figure 1**, when $\alpha = 0.05$ in each individual test. The five rejection regions imply that each test has optimal power against different alternatives. The Wei-Lachin SUM test rejects $(\widehat{\beta}_1, \widehat{\beta}_2)$ outside a straight line with slope –1 which represents a constant sum. The rejection region of the Two-Step test can be viewed as removing two sides from the rejection region of the SUM test. The MAX test and ALRT reject points with a large positive value in at least one dimension. The rejection region of the IUT eliminates points with negative or close to zero values in any endpoint compared to the rejection region of ALRT.

3. SIMULATION STUDIES

We simulate binary data following a logistic model to illustrate the global tests. Two different scenarios are examined – correlated multiple outcomes and independent stratified data. For correlated outcomes, each subject i has two endpoints. The independent data are from two strata. Two independent covariates are generated: a binomial variable X_{1ij} with equal probability to be zero or one and a normal variable X_{2ij} with mean 0 and standard deviation 5. The outcomes Y_{ij} follow Bernoulli distribution specified by $logit\{p(Y_{ij} = 1)\} = \eta_j + \beta_j X_{1ij} + \theta X_{2ij}$. Note the effects of the treatment X_{1ij} is reflected by two endpoint-specific regression coefficients, β_1 and β_2. Correlated binary outcomes are generated following Park, Park, and Shin method (19). The intercepts for endpoint 1 and 2 are $\eta_1 = 0.5$, $\eta_2 = 0.2$, and the coefficient for X_{2ij} is $\theta = 0.1$ for both endpoints. In simulating the independent binary data, $\theta = 0.02$. In case of unequal sample sizes in the two endpoints, observations in the endpoint with less subjects are missing completely at random. Maximum likelihood estimator for β_1 and β_2, as well as the covariance matrix $\sum(\widehat{\beta}_1, \widehat{\beta}_2)$ are calculated through generalized estimating equations (20). One-sided alternatives H_a: $(\beta_1 > 0) \cap (\beta_2 > 0)$ are tested. Test statistics are calculated using $\widehat{\beta}_1, \widehat{\beta}_2$. Each setup is repeated 1000 times. In each iteration, all the test statistics are calculated using the same dataset. We examine and compare the powers and Type I error rates of all five tests for different true values (**Table 3**), different levels of correlations (**Table 4**), and different sample sizes at each endpoint (**Table 5**).

The powers and Type I error rates for different (β_1, β_2) values are listed in **Table 3**. The correlation between Y_{i1} and Y_{i2} is set to be 0.4 and each endpoint has 100 observations. All tests except the IUT maintain the Type I error rates close to the nominal level 0.05. Without prior knowledge of the indifference range, we set the most restrictive indifference range where $\varepsilon_j = 0$ for every endpoint which is equivalent to requiring all treatment effects to be positive, leading to low overall type I error rate and power. IUT tends to be more conservative than other methods because FDA is more concerned with false positives and only approves new treatment when there is significant evidence supporting its superiority. The procedures can be divided into two groups according to how the power changes when the difference between β_1 and β_2 gets larger. The first group includes Wei-Lachin SUM and Two-Step. They are more powerful than the other group when $\beta_1 = \beta_2$, but sensitive to non-homogeneous treatment effects. The power of the Wei-Lachin SUM test drops from 52 to 21% when β_2 drops from 0.6 to 0 while β_1 remains 0.6. The decreasing trend is even more obvious with the Two-Step. The second group includes ALRT and the MAX test. They are robust to non-homogeneous treatment effects.

In **Table 4**, 100 correlated pairs (Y_{i1}, Y_{i2}) are generated with various correlation coefficients. All the methods incorporate information from both endpoints. When two outcomes are highly correlated, the treatment effects estimated from both endpoints, $\widehat{\beta}_1$ and $\widehat{\beta}_2$, tend to be similar and provide less information compared to the independent case, hence lower power. However, the IUT has a reversed pattern because with higher positive correlation, the non-inferiority tests on the two endpoints tend to agree more, leading to higher overall rejection rates.

Table 5 lists the different performance of the tests with unequal sample sizes for the two endpoints. When sample sizes are not balanced between the two endpoints, most tests have reduced power because the test statistics combine information from all endpoints and a large variance in one endpoint leads to large variance of the overall test statistic. ALRT is robust to unequal sample sizes. If

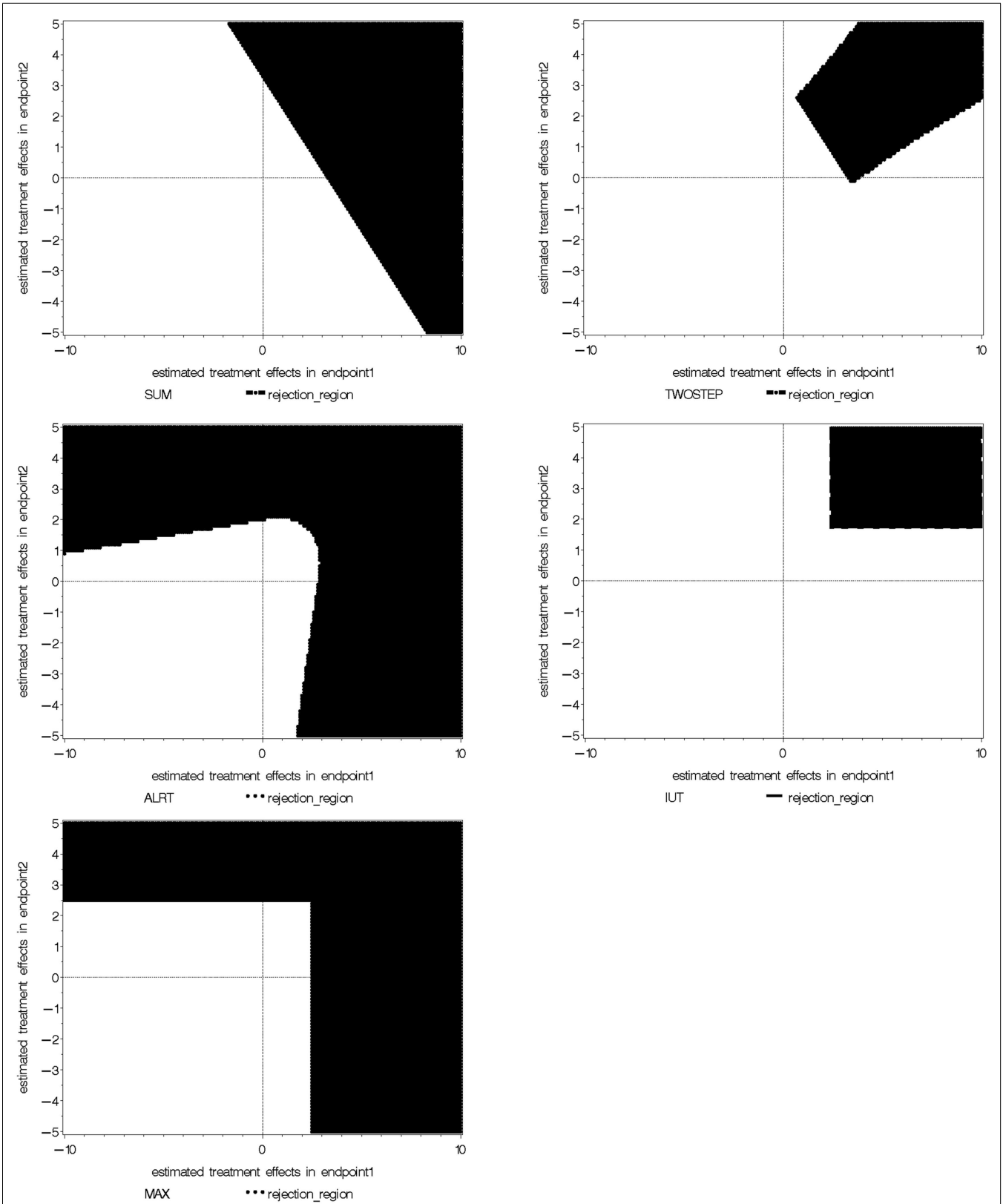

FIGURE 1 | Comparison of rejection regions of five global tests.

Table 3 | Simulation results: size and power (%) with different true value positions.

True values		ρ	Testing procedure				
β_1	β_2		SUM	Two-step	IUT	ALRT	MAX
0	0	0.4	4.9	4.7	0.7	4.4	4.5
0.3	0.3	0.4	21	20	5	18	18
0.6	0.6	0.4	52	50	23	48	46
0.6	0.3	0.4	35	33	11	35	36
0.6	0	0.4	21	17	3	32	33
0.6	−0.3	0.4	11	7	1	33	29

The Type I error 4.7% for Two-Step refers to cases rejected in Step Two. The Type I error for the three tests in IUT are 4.3%, 5.1%, and 4.4%, respectively. The intersection of the three rejection regions gives the overall Type I error rate.

Table 4 | Simulation results: power (%) under different correlation between outcomes.

True values		ρ	Testing procedure				
β_1	β_2		SUM	Two-step	IUT	ALRT	MAX
0.6	0.6	0	63	61	18	58	52
0.6	0.6	0.4	52	50	23	48	46
0.6	0.6	0.8	42	40	27	37	41
0.6	0.3	0	44	42	7	40	38
0.6	0.3	0.4	35	33	11	35	36
0.6	0.3	0.8	26	24	12	28	30

ρ Represents the correlation coefficient between the outcomes, not the correlation between the estimated regression coefficient.

treatment effects are equal in both endpoints ($\beta_1 = \beta_2 = 0.6$), the power of ALRT does not change with the distribution of samples into each two endpoints as long as the total sample size remains the same. When the treatment effects differ in the two endpoint ($\beta_1 = 0.6$, $\beta_2 = 0.3$), the power of ALRT could either increase or decrease depending on which endpoint has more subjects. If the endpoint with a larger treatment effects has a larger sample size, ALRT has higher power. If the endpoint with a smaller treatment effect gets more samples, the power decreases.

4. CONTROLLING FWER/FDR IN GENOMIC STUDIES

4.1. STEPWISE PROCEDURES AND FDR

Stepwise procedures are classified into one-step procedures and multi-step procedures. One-step procedures set a uniform threshold for all the unadjusted *P*-values while multi-step procedures set different thresholds for different hypotheses depending on the order of the unadjusted *P*-values. Multi-step procedures can be carried out step-down or step-up (21, 22). In step-down procedures, the hypothesis with the smallest *P*-value is tested first. And as long as one hypothesis fails to be rejected, all the hypotheses with larger unadjusted *P*-values will fail to be rejected. On the contrary, step-up procedures start from the largest unadjusted *P*-value and reject all smaller unadjusted *P*-values after the first one is rejected.

Table 5 | Simulation results: power (%) with different sample sizes.

True values		Sample size		ρ	Testing procedure				
β_1	β_2	n_1	n_2		SUM	Two-step	IUT	ALRT	MAX
0.6	0.6	100	100	0.4	52	50	23	48	46
0.6	0.6	50	150	0.4	41	37	15	49	30
0.6	0.6	25	175	0.4	30	26	5	51	16
0.6	0.3	100	100	0.4	35	33	11	35	36
0.6	0.3	50	150	0.4	27	26	7	23	21
0.6	0.3	25	175	0.4	18	18	3	23	9
0.6	0.6	100	100	0	63	61	18	58	52
0.6	0.6	50	150	0	55	50	14	55	37
0.6	0.6	25	175	0	36	34	9	54	17
0.6	0.3	100	100	0	44	42	7	40	38
0.6	0.3	150	50	0	37	33	7	47	24
0.6	0.3	175	25	0	24	22	5	48	10

In this manuscript, the FWER is preserved at nominal level in a strong sense, that is, FWER is no larger than the nominal level for testing any subset of the hypotheses set. Given the large number of hypotheses, researchers are often more interested in a more interpretable quantity, the FDR (23). FDR is the rate that the rejected or significant features are truly null. The numbers of true and false positives can be calculated directly from FDR. FDR helps to avoid a flood of false positives when most of the hypotheses are truly null or missing out significant features when the number of true alternative hypotheses is large. FDR can be estimated as

$$FDR = \frac{\pi_0 mt}{\sum_{j=1}^{J} I(P_j \leq t)},$$

where m is the total number of hypotheses being tested, π_0 is the percent of true null among them, I is the indicator for a true statement in the bracket, and t is the cutoff value of p-values to call a feature significant. Although π_0 is unknown, it can be estimated from the distribution of *P*-values. Benjamini and Hochberg (24) developed a step-up procedure to control FDR at level q^*. For ordered unadjusted *P*-values $P_{(1)}, P_{(2)}, \ldots, P_{(J)}$, we reject the first j hypotheses with the smallest j *P*-values if $P_{(j)} \leq \frac{jq^*}{J}$.

4.2. RESAMPLING METHOD

Westfall and Young (25) and Troendle (26) developed resampling procedures which simulate the joint distributions of the *P*-values under the null while maintaining their correlation structure. The procedure starts with bootstrap or permutation under the null from the original sample. Then hypothesis-specific pivotal test statistics and the corresponding *P*-values are calculated on the simulated data. The steps are repeated a large number of times to achieve an empirical distribution of $(P_1, P_2, \ldots, P_J)$ under the null which maintains the correlation structure. The unadjusted *P*-values for the jth hypothesis is the percent of times the jth imputed test statistic is larger than or equal to the jth test statistic from the original data. Step-down or step-up procedures can be carried out on the unadjusted *P*-values based on resampling. There is a resampling option in SAS "multtest" procedure for several tests including

Table 6 | Real data analysis: association between log(GFR) and 14 SNPs.

SNP	Minor	Raw	Bonferroni	Sidak	Hochberg	FDR
rs307806	A	0.01071	0.1499	0.1399	0.1499	0.1499
rs2279622	T	0.03383	0.4736	0.3823	0.4398	0.1939
rs4693614	G	0.06319	0.8846	0.5990	0.6924	0.1939
rs11715496	A	0.06702	0.9382	0.6214	0.6924	0.1939
rs8042694	G	0.06924	0.9693	0.6338	0.6924	0.1939
rs2259458	T	0.22639	1.0000	0.9725	0.9791	0.4245
rs3824935	T	0.23555	1.0000	0.9767	0.9791	0.4245
rs2027440	C	0.24256	1.0000	0.9795	0.9791	0.4245
rs2276768	T	0.30994	1.0000	0.9944	0.9791	0.4821
rs10497435	C	0.44626	1.0000	0.9997	0.9791	0.6248
rs3814995	T	0.52058	1.0000	1.0000	0.9791	0.6626
rs2705897	T	0.61445	1.0000	1.0000	0.9791	0.7169
rs7844961	T	0.73593	1.0000	1.0000	0.9791	0.7925
rs4900312	A	0.97914	1.0000	1.0000	0.9791	0.9791

the two-sample t-test, Cochran-Armitage test and Fisher's exact test. However, this procedure does not allow covariate adjustments and can not be used in multiple comparisons in regressions.

4.3. BONFERRONI ADJUSTMENT

The Bonferroni adjustment is a one-step procedure which rejects the jth null hypothesis H_{0j} when the p-value in testing the jth hypothesis $P_j \leq \frac{\alpha}{J}$. The FWER in the Bonferroni procedure is conserved at α level because

$$\begin{aligned} Pr(\text{Reject any } H_{0j}|H_0) &= Pr\left\{\bigcup_{j=1}^{J}(P_j \leq \frac{\alpha}{J})|H_0\right\} \\ &\leq \sum_{j=1}^{J} Pr(P_j \leq \frac{\alpha}{J}|H_0) \\ &= \sum_{j=1}^{J} \frac{\alpha}{J} \\ &= \alpha, \end{aligned}$$

where $\cup$ denotes union. Alternatively, researchers may compute adjusted P-values as $P_j^* = P_j \times J$ and compare P_j^* to the nominal level α. The Bonferroni adjustment is computationally straight forward because the threshold for significant P-values in each hypothesis is just the FWER divided by the number of hypotheses. However, Bonferroni procedure is conservative with low power.

Wiens (27) and Huque and Alosh (28) modified the Bonferroni procedure with fixed testing sequence procedure. It allocates the overall Type I error rate sequentially and controls FWER at the nominal level. Let the sequence of hypotheses be $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(J)}$. Assign type I error rate α_j to each of the null hypothesis such that $\sum_{j=1}^{J} \alpha_j = \alpha$. Furthermore, if the first hypothesis is not rejected, its portion of the type I error α_1 will be passed onto the second hypothesis. That is, the type I error rate for the second hypothesis becomes $\alpha_1 + \alpha_2$ conditional on that $H_0^{(1)}$ fails to be rejected. On the contrary, if $H_0^{(1)}$ is rejected, the type I error rate of $H_0^{(2)}$ remains α_2. In summary, the type I error rates of unrejected hypotheses accumulate and are passed onto the next hypotheses until a hypothesis is rejected or the last hypothesis $H_0^{(J)}$.

4.4. HOLM, SIDAK, AND SIMES PROCEDURES

Holm method is a step-down procedure (29). First, it ranks all the observed p-values from smallest to largest $P_{(1)}, P_{(2)}, \ldots, P_{(J)}$ Compare each P_j to $\frac{\alpha}{J+1-j}$ starting from the smallest $P_{(1)}$. Let the first occurrence of $P_{(j)} > \frac{\alpha}{J+1-j}$ be the kth ordered p-value. Then hypotheses corresponding to the first $k-1$ p-values $P_{(1)}, \ldots, P_{(k-1)}$ will be rejected and the hypotheses from the kth one on corresponding to $P_{(k)}, \ldots, P_{(J)}$ will not be rejected. Alternatively, researchers can also compute the Holm's adjusted p-values and compare them to α. The adjusted p-values is based on the ordered p-values $P_{(1)}, P_{(2)}, \ldots, P_{(J)}$ and $P_{(j)}^* = \max_{i \leq j}\{(J-i+1)P_{(i)}^* \wedge 1\}$ where $\wedge$ denotes taking the minimum. The adjusted p-values are capped at 1 by taking the minimum of $(J-i+1)P_{(i)}^*$ and 1. Besides, the jth adjusted p-value is the maximum in the first j values, resulting in non-decreasing sequence of adjusted P-values.

The Sidak (30) correction assumes that the J test statistics are mutually independent and replaces the element-wise p-value cut-off α/J by $1-(1-\alpha)^{\frac{1}{J}}$. It is less conservative than the Bonferroni correction because $1-(1-\alpha)^{\frac{1}{J}} \geq \frac{\alpha}{J}$ for $n \geq 1$. Another set of thresholds combining the Holm threshold and Sidak correction, $1-(1-\alpha)^{\frac{1}{J}}, 1-(1-\alpha)^{\frac{1}{J-1}}, \ldots, 1-(1-\alpha)^{\frac{1}{1}}$, also maintains FWER at α.

Simes (31) procedure is also a step down procedure that rejects H_{0j} when $P_{(j)} \leq \frac{j\alpha}{J}$. Here $P_{(1)}, P_{(2)}, \ldots, P_{(J)}$ are the ordered P-values from smallest to largest. Hochberg and Liberman (32) extended the Simes procedure by allocating different weights to the P-values depending on prior information on each hypothesis.

5. A REAL CASE: GENOMIC STUDIES BASED ON A CLINICAL TRIAL

We illustrate the stepwise procedures using the Genome Wide Association Study (GWAS) from the Diabetes Control and Complications Trial (DCCT) and Epidemiology of Diabetes Intervention and Complication (EDIC) trial. DCCT and EDIC are two clinical trials based on the same type 1 diabetes cohort in different time periods. The survival rate and life expectancy of type 1 diabetic patients have been improved greatly in recent years. However, chronic hyperglycemia status leads to deleterious changes in blood vessels. Cardiovascular diseases and microvascular complications are major threats to the long-term quality of life of type 1 diabetic patients. This study focuses on microvascular complications among type 1 diabetic patients. In EDIC, 1441 Type 1 diabetic patients enrolled from 1983 to 1989. They were randomized to either the intensive or conventional therapies, where participants in the intensive group monitored and regulated their blood glucose level constantly. DCCT ended in 1993 when significant reduction in the risk of microvascluar complications was found in the intensive therapy group (33). Of the 1441 DCCT participants, 1394

continued to the EDIC trial, where everyone receives the intensive therapy.

The abnormalities in the capillaries lead to symptoms in different parts of the body – nephropathy, retinopathy, and neuropathy. The goal of this analysis is to validate fourteen SNPs associated with severe nephropathy and persistent microalbuminuria in Al-Kateb et al. (34). Urine glomerular filtration rates (GFR), which is an important clinical index of diabetic nephropathy, have been recorded annually in the DCCT/EDIC cohort. Log-transformed GFR values are employed as our main outcome. Linear regressions of the last GFR observation versus each of the fourteen SNPs are fitted, adjusting for age at randomization, gender and duration of diabetes at enrollment, stratified by the treatment group. Different SNP coefficients are assumed in the intensive and conventional treatment groups because patients under the two treatments were in quite different biophysical and metabolic statuses and the intensive control of the glucose level might suppress or activate SNP effects. Among the global tests for SNP effects in different strata, the SUM test is employed as the effects for each SNP are expected to be in the same direction across the two strata and the SUM test is the maxmin test under such conditions.

Fourteen raw *P*-values are generated from the SUM tests, one for each SNP. To maintain the family wise Type I error rate or false discovery rate, four different stepwise procedures are performed – Bonferroni, Sidak, Hochberg, and FDR. All four procedures are directly available in SAS package "multtest." *P*-values of various procedures are listed in **Table 6**. We can see that although some raw *P*-values are <0.05, none of the adjusted *P*-values remain significant. That is, after FWER is controlled, the seemingly significant results are not actually significant any more. Among the procedures controlling FWER, the Sidak and Hochberg procedures give smaller adjusted *P*-values and therefore are more powerful than the Bonferroni adjustment. Although researchers usually require the FWER no larger than 0.05, they might set higher cutoff value of FDR depending on the context of the research problem.

6. DISCUSSION

This manuscript reviews methods for the multiple hypothesis testing problem. Five global tests widely used in clinical trials are reviewed: SUM test, Two-Step test, ALRT, IUT, and the MAX Test. The plots of the rejection regions illustrate the different alternatives to which the tests are directed. The SUM and Two-Step tests are powerful for alternatives with homogeneous effects. Two-Step test can be viewed as a modification of the SUM test that incorporates information on how different the treatment effects are and thus more sensitive to non-homogeneous treatment effects. ALRT is robust to not only unequal treatment effects but also unequal sample sizes from the endpoints. MAX test is also robust for non-homogeneous treatment effects. IUT provides information about the overall superiority and individual non-inferiority. In genomic studies, specific conclusions on individual hypotheses are desired and stepwise procedures are commonly used to control FWER or FDR. The Westfall and Young's resampling method generates the joint distribution of *P*-values under the null and maintains the correlation structure between them. A selected SNP dataset from a clinical trial is used to illustrate the stepwise procedures. Finally, among the hundreds of papers on multiple hypothesis testing topic, only a selected few commonly used multiple hypothesis testing adjustment methods are reviewed here. Our goal is to introduce the classical methods and present the ideas behind them. They serve as the basis on which researchers may choose and develop their own method with careful consideration of the particular research setup and clinical questions.

ACKNOWLEDGMENTS

The author thanks Professor John Lachin and Professor Joseph Gastwirth for many helpful discussions. We also wish to thank the EDIC research group for access to the GFR and GWAS data. This publication was supported by Award Numbers UL1TR000075 and KL2TR000076 from the NIH National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. This study was approved by the IRB office at George Washington University.

REFERENCES

1. Logan BR, Tamhane AC. Combined global and marginal tests to compare two treatments on multiple endpoints. *Biomed J* (2001) **43**:591–604. doi:10.1002/1521-4036(200109)43:5<591::AID-BIMJ591>3.0.CO;2-F
2. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* (1984) **40**:1079–87. doi:10.2307/2531158
3. Tang DI, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* (1993) **49**:23–30. doi:10.2307/2532599
4. Lauter J, Glimm E, Kropf S. New multivariate tests for data with an inherent structure. *Biomed J* (1996) **37**:5–23.
5. Bregenzer T, Lehmacher W. Directional tests for the analysis of clinical trials with multiple endpoints allowing for incomplete data. *Biomed J* (1998) **40**:911–28. doi:10.1002/(SICI)1521-4036(199812)40:8<911::AID-BIMJ911>3.0.CO;2-W
6. Gastwirth JL. The use of maximum efficiency robust tests in combining contingency tables and survival analysis. *J Am Stat Assoc* (1985) **80**:380–4. doi:10.1080/01621459.1985.10478127
7. Wei LJ, Lachin JM. Two-sample asymptotically distribution free tests for incomplete multivariate observations. *J Am Stat Assoc* (1984) **79**:653–61. doi:10.1080/01621459.1984.10478093
8. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* (1987) **43**(2):487–98. doi:10.2307/2531989
9. Frick H. A maximum linear test of normal means and its application to Lachin's data. *Comm Stat Theor Meth* (1994) **23**(4):1021–9. doi:10.1080/03610929408831302
10. Breslow N, Day NE. The analysis of case-control studies. *Statistical Methods in Cancer Research* (Vol. 1), Lyon: France, IARC Scientific Publications (1980).
11. Legler JM, Lefkopoulou M, Ryan LM. Efficiency and power of tests for multiple binary outcomes. *J Am Stat Assoc* (1995) **90**:680–93. doi:10.1080/01621459.1995.10476562
12. Lachin JM, Wei LJ. Estimators and tests in the analysis of multiple nonindependent 2x2 tables with partially missing observations. *Biometrics* (1988) **44**:513–28. doi:10.2307/2531864
13. Follmann D. A simple multivariate test for one-sided alternatives. *J Am Stat Assoc* (1996) **91**:854–61. doi:10.1080/01621459.1996.10476953
14. Tang DI, Gnecco C, Geller NL. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* (1989) **76**:577–83. doi:10.1002/bimj.200900203
15. Tarone RE. On the distribution of the maximum of the log-rank statistics and the modified Wilcoxon statistics. *Biometrics* (1981) **37**:79–85. doi:10.2307/2530524

16. Bloch DA, Lai TL, Tubert-Britter P. One-sided tests in clinical trials with multiple endpoints. *Biometrics* (2001) **57**:1039–47. doi:10.1111/j.0006-341X.2001.01039.x
17. Perlman MD, Wu L. A note on one-sided tests with multiple endpoints. *Biometrics* (2004) **60**:276–80. doi:10.1111/j.0006-341X.2004.00159.x
18. Chen JJ, Wang SJ. Testing for treatment effects on subsets of endpoints. *Biomed J* (2002) **44**:541–57. doi:10.1002/1521-4036(200207)44:5<541::AID-BIMJ541>3.0.CO;2-0
19. Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. *Am Stat* (1996) **50**:306–10. doi:10.1080/00031305.1996.10473557
20. Zeger SL, Liang KY. Longitudinal data analysis using generalized linear models. *Biometrika* (1986) **73**:13–22. doi:10.1093/biomet/73.1.13
21. Dunnett CW, Tamhane AC. A step-up multiple test procedure. *J Am Stat Assoc* (1982) **87**:162–70. doi:10.1080/01621459.1992.10475188
22. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* (1991) **47**:511–21. doi:10.2307/2532142
23. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* (2003) **100**(16):9440–5. doi:10.1073/pnas.1530509100
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* (1995) **57**:289–300.
25. Westfall PH, Young SS. *Resampling-Based Multiple Testing.* New York: Wiley (1993).
26. Troendle JF. A stepwise resampling method of multiple hypothesis testing. *J Am Stat Assoc* (1995) **90**:370–8. doi:10.1080/01621459.1995.10476522
27. Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharm Stat* (2003) **2**:211–5. doi:10.1002/pst.64
28. Huque MF, Alosh M. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *J Stat Plan Inference* (2008) **138**:321–35. doi:10.1016/j.jspi.2007.06.009
29. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* (1979) **6**(2):65–70.
30. Sidak ZK. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc* (1967) **62**(318):626–33. doi:10.1080/01621459.1967.10482935
31. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* (1986) **73**:751–4. doi:10.1093/biomet/73.3.751
32. Hochberg Y, Liberman U. An extended Simes test. *Stat Probab Lett* (1994) **21**:101–5. doi:10.1016/0167-7152(94)90216-X
33. The DCCT Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* (1993) **329**:977–86. doi:10.1056/NEJM199309303291401
34. Al-Kateb H, Boright AP, Mirea L, Xie X, Sutradhar R, Mowjoodi A, et al. Multiple superoxide dismutase 1/splicing factor serine alanine 15 variants are associated with the development and progression of diabetes nephropathy: the diabetes control and complications trial/epidemiology of diabetes interventions and complications genetic study. *Diabetes* (2008) **57**:218–28. doi:10.2337/db07-1059

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 November 2013; accepted: 15 November 2013; published online: 09 December 2013.
Citation: Pan Q (2013) Multiple hypotheses testing procedures in clinical trials and genomic studies. Front. Public Health ***1****:63. doi: 10.3389/fpubh.2013.00063*
This article was submitted to Epidemiology, a section of the journal Frontiers in Public Health.