Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 14

Quantile-Function Based Null Distribution in Resampling Based Multiple Testing

Mark J. van der Laan* Alan E. Hubbard†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

Quantile-Function Based Null Distribution in Resampling Based Multiple Testing

Mark J. van der Laan and Alan E. Hubbard

Abstract

Simultaneously testing a collection of null hypotheses about a data generating distribution based on a sample of independent and identically distributed observations is a fundamental and important statistical problem involving many applications. Methods based on marginal null distributions (i.e., marginal p-values) are attractive since the marginal p-values can be based on a user supplied choice of marginal null distributions and they are computationally trivial, but they, by necessity, are known to either be conservative or to rely on assumptions about the dependence structure between the test-statistics. Re-sampling based multiple testing (Westfall and Young, 1993) involves sampling from a joint null distribution of the test-statistics, and controlling (possibly in a, for example, step-down fashion) the user supplied type-I error rate under this joint null distribution for the test-statistics. A generally asymptotically valid null distribution avoiding the need for the subset pivotality condition for the vector of test-statistics was proposed in Pollard, van der Laan (2003) for null hypotheses about general real valued parameters. This null distribution was generalized in Dudoit, vanderLaan, Pollard (2004) to general null hypotheses and test-statistics. In ongoing recent work van der Laan, Hubbard (2005), we propose a new generally asymptotically valid null distribution for the test-statistics and a corresponding bootstrap estimate, whose marginal distributions are user supplied, and can thus be set equal to the (most powerful) marginal null distributions one would use in univariate testing to obtain a p-value. Previous proposed null distributions either relied on a restrictive subset pivotality condition (Westfall and Young) or did not guarantee this latter property (Dudoit, vanderLaan, Pollard, 2004). It is argued and illustrated that the resulting new re-sampling based multiple testing methods provide more accurate control of the wished Type-I error in finite samples and are more powerful. We establish formal results and investigate the practical performance of this methodology in a simulation and data analysis.

KEYWORDS: asymptotic control, bootstrap, multiple testing, null distribution, quantile-quantile function, type I error rate

1 Introduction

Recent technological developments in biological research, for instance genomics and proteomics, have created new statistical challenges by providing simultaneously thousands of biological measurements (e.g., gene expressions) on the same experimental unit. Typically, the collection of these measurements is made to determine, for example, which genes of the thousands of candidates are associated with some other, often phenotypic, characteristic (e.g., disease status). This has led to the problem of properly accounting for simultaneously testing a large number of null hypotheses when making inferences about the tests for which the null is rejected. Multiple testing is a subfield of statistics concerned with proposing procedures involving a rejection or acceptance decision for each null hypothesis. Multiple testing procedures (MTP's) are used to control various parameters of either the distribution of the number of false rejections or the proportion of false rejections, and these are often referred to as different varieties of Type-I error rates. In addition, among such procedures controlling a particular Type-I error rate (for a desired type I error rate of α, the procedure guarantees the error rate is $\leq \alpha$), one aims to find a procedure which has maximal power in the sense that it finds more of the true positives than competing procedures. Parallelling the introduction of new high-throughput technologies that demand adjustment for multiple comparisons, there has been a growing list of proposals for MTP's (e.g., Efron et al. (2001), Genovese and Wasserman (2003), Lehmann and Romano (2005)).

Methods based on marginal null distributions, or equivalently, marginal p-values, are known to either be conservative (by having to be valid under all possible joint distributions for the test-statistics) or have to rely on assumptions about the dependence structure between the test-statistics. Resampling based multiple testing involves estimating a joint null distribution of the test-statistics, and controlling the user supplied type-I error rate under this data dependent joint null distribution.

In re-sampling based multiple testing it has been common practice to enforce the null distribution to correspond with a data generating distribution which satisfies the overall null hypothesis (Westfall and Young (1993)). This typically results in a distribution which does not correctly specify the dependence structure (e.g., covariance matrix) of the true distribution of the test-statistics, and thereby does not guarantee the wished type-I error control under the true distribution: i.e., it relies on the so called subset pivotality

condition introduced and discussed in Westfall and Young (1993).

In order to avoid this restrictive subset-pivotality condition, a generally valid null distribution was originally proposed in Pollard and van der Laan (2003) for tests concerning real valued parameters, and it was generalized to general hypotheses and general test-statistics in a subsequent article (Dudoit et al. (2004)). They choose as general null distribution, the null-value shifted distribution of the test-statistics (e.g., centered t-statistic), which conserves the dependence structure of the test-statistics, and thereby guarantees that the number of false rejections under the true distribution is dominated by the number of false rejections under their proposed null distribution. They showed that the latter null distribution is naturally estimated with the model-based or nonparametric bootstrap, which involves sampling from an estimate of the true distribution of the data. If the null-value shifted marginal distribution of a test-statistic has a larger variance than one would have under a true null hypothesis, then they proposed to also scale the variance of the marginal distribution to the null value for the variance, which improves the power of the corresponding multiple testing procedure.

Typically, if the null hypothesis is true, then the marginal distribution of a test-statistic is known. The null-value shifting and scaling of the marginal distribution of the test-statistics, as proposed in Dudoit et al. (2004), guarantees that the obtained marginal distribution has the mean and variance of this known marginal null distribution, but it does not guarantee that the whole marginal distribution is equal to this marginal null distribution. In particular, this proposed joint null distribution does not generalize the univariate null distribution one would use in univariate testing, and thereby does not necessarily imply the most powerful univariate testing procedure. Given that the previously proposed joint null distribution does not necessarily have the optimal marginal null distributions, one can expect that it should be possible to improve the power of the multiple testing procedure.

This motivates us to construct a new valid null distribution for the test-statistics whose marginal distributions are exactly equal to the wished *user-supplied* marginal null distributions one would use in univariate testing. As a consequence, marginal p-values under this newly proposed joint null distribution are now (can be chosen to be) equal to the marginal p-values one would use in univariate testing. As a consequence, the so called Type-I error adjusted p-values of a multiple testing procedure based on this joint null distribution, which in addition to getting the marginal null distribution of the test-statistic right, also capitalizes on the dependence among the

DOI: 10.2202/1544-6115.1199

test-statistics, are guaranteed to be smaller than the adjusted-p-values of the analogue of this multiple testing procedure only using the marginal null distributions.

Any of the re-sampling based multiple testing procedures controlling a particular type-I error rate proposed in the literature can now be applied to this new joint null distribution. In particular, we can apply our new joint null distribution to 1) the step-down re-sampling based multiple testing procedures controlling family-wise error (FWE; Westfall and Young (1993),van der Laan et al. (2004c)), 2) the single step re-sampling based multiple testing procedures controlling generalized FWE (Dudoit et al. (2004), 3) the re-sampling based augmentation methods controlling the Tail probability of the proportion of false positive among the rejections at $q \in (0,1)$ (TPPFP(q)) (van der Laan et al. (2004b)), and 4) the Empirical Bayes resampling based multiple testing method controlling TPPFP(q) (van der Laan et al. (2004a)). This results in a new set of single step and step down re-sampling based multiple testing procedures MTP's asymptotically controlling a user supplied Type-I error rate. By utilizing the best of marginal p-value methods and re-sampling based methods, this class of procedures, under general data-generating distributions, should be as powerful or more powerful than any existing MTP. In addition, we have found that the existing re-sampling based MTP's can suffer from inaccurate control unless enormous number of bootstrap samples are performed. By being able to specify the marginal null distributions in our proposed joint null distribution, it has been our practical experience that one needs a much smaller number of bootstrap samples to achieve the wished performance. This makes our new re-sampling based MTP's not only more powerful, but also more practical than previous re-sampling based multiple testing procedures.

1.1 Organization.

In Section 2, we first present the statistical framework of multiple testing, and present the new joint null distribution and its bootstrap estimate. Unfortunately, this section relies heavily on notation. To aid the reader, we provide a glossary of notation in the appendix. We will stay consistent with notation from previous papers (Pollard and van der Laan (2003), Dudoit et al. (2004), van der Laan et al. (2004c), van der Laan et al. (2004d)). We formally establish that the sub-distribution corresponding with the true null hypotheses of this new joint null distribution asymptotically dominates the corresponding

true sub-distribution of the test-statistics. In Section 3, we provide a general theorem establishing that a single step multiple testing procedure based on this null distribution asymptotically controls the wished type-I error rate. For additional formal results, stating asymptotic control of the wished Type-I error rate for single step and step-down multiple testing procedures based on this joint null distribution, we refer to general theorems in Dudoit et al. (2004) and van der Laan et al. (2004c) which can be applied to our new joint null distribution. In Section 4 we work out our proposed multiple testing methodology for two general classes of multiple testing problems. First, as in Pollard and van der Laan (2003), we consider null hypotheses stating that a real valued parameter is smaller or equal than a hypothesized value, where we allow any kind of real valued parameter that can be estimated at a root-n-rate. For example, in genetic studies one might wish to test the null hypothesis that a particular single nucleotide polymorphism (SNP) has no effect on the mean of a phenotypic outcome for each SNP among a set of SNP's. In the second subsection of Section 4 we consider null hypotheses stating that K real valued parameters are equal. In Section 5 we provide a simulation study in order to investigate the practical performance of our bootstrap null distribution in re-sampling based multiple testing; in Section 6 we perform an analysis on SNP/cancer data and end with a discussion.

2 Multiple testing framework and the new joint null distribution

Statistical framework for multiple testing: Let $X_1, \ldots, X_n$ be n i.i.d. observations of $X \sim P$, where P is known to be an element of a model $\mathcal{M}$. Consider a collection of null hypotheses of the form $H_{0j} : P \in \mathcal{M}_j$ for subsets $\mathcal{M}_j \subset \mathcal{M}$, $j = 1, \ldots, m$. Throughout this paper we will let $T_n = (T_n(1), \ldots, T_n(m))$ be a vector of test-statistics with unknown distribution $Q_n(P)$ corresponding with this set of null hypotheses $H_{01}, \ldots, H_{0m}$ such that large values of $T_n(j)$ provide statistical evidence that the null hypothesis H_{0j} is false, and n indicates the sample size. Here T_n is a test-statistic vector based on a sample of n i.i.d. $X_1, \ldots, X_n$ with a common distribution P so that the distribution $Q_n = Q_n(P)$ of T_n is identified by the data generating distribution P. Let P_n denote the empirical distribution of $X_1, \ldots, X_n$. In order to stress the dependence of T_n on the data, we will also use the no-

DOI: 10.2202/1544-6115.1199

tation $T_n = T_n(P_n)$. Let $\mathcal{S}_0 \equiv \{j : H_{0j} \text{ is true}\}$ denote the set of true null hypotheses.

A multiple testing procedure is a random subset $\mathcal{S}_n \subset \{1, \ldots, m\}$ indicating the set of null hypotheses which are claimed to be false. Given a multiple-testing procedure one can define the number of false positives $V_n =\mid \mathcal{S}_n \cap \mathcal{S}_0 \mid$ under the true data generating distribution P. Consider a type-I error rate $\theta(F_{V_n}) \in \mathbb{R}$, where F_{V_n} denotes the cumulative distribution function of the number of false rejections, V_n, of the given multiple testing procedure $\mathcal{S}_n$. The multiple testing literature is concerned with providing powerful multiple testing procedures controlling $\theta(F_{V_n}) \leq \alpha$.

Consider a multiple testing procedure $S(T_n, Q, \alpha) = \{j : T_n(j) > c_j(Q, \alpha)\}$ based on a cut-off rule $c(Q, \alpha)$ aiming to control a particular Type-I error rate at level α, by controlling this Type-I error under a null distribution Q. For example, a single-step multiple testing procedure corresponds with setting the cut-off vector $c(Q, \alpha)$ so that $\theta(F_{R_n(c(Q,\alpha)|Q)}) = \alpha$, where $R_n(c \mid Q) = \sum_{j=1}^m I(Z(j) > c_j)$, $Z \sim Q$, is the number of rejections under the null distribution Q. Given such a multiple-testing procedure one can define the number of false positives V_n as $V_n = \sum_{j=1}^m I(T_n(j) > c_j(Q, \alpha), j \in \mathcal{S}_0)$, where $T_n \sim Q_n(P)$ is distributed as the actual vector of test-statistics. Typically, Q is an estimated null distribution so that it is also random through P_n.

In this section we propose a choice of null distribution (i.e., Q) which guarantees the wished asymptotic control of $\theta(F_{V_n})$ at level α. It will be assumed that this real valued parameter $F \to \theta(F)$ on cumulative distribution functions on $\{1, \ldots, m\}$ satisfies the monotonicity assumption (AMI) and the continuity assumption (ACI), as defined in Dudoit et al. (2004): that is, if $F \leq G$, then $\theta(F) \geq \theta(G)$, and if $F_n - G_n$ converges to zero for $n \to \infty$, then $\theta(F_n) - \theta(G_n)$ converges to zero for $n \to \infty$. As shown in the latter paper, these two conditions hold for all well known Type-I error rates (e.g, $\theta(F) = 1 - F(0)$ (FWE)).

Definition of finite sample joint null distribution: Let Q_{0j} be a marginal user-supplied null distribution, $j = 1, \ldots, m$, so that for $j \in \mathcal{S}_0$ and all x

$$\lim \inf_{n \to \infty} Q_{0j}^{-1} Q_{nj}(x) \geq x, \tag{1}$$

where Q_{nj} is the j-th marginal distribution of the true distribution $Q_n = Q_n(P)$ of the test-statistic vector T_n. That is, for $j \in \mathcal{S}_0$, the marginal distribution Q_{nj} of $T_n(j)$ is asymptotically dominated by this null distribution

Q_{0j}.

We note that the finite sample validity of the proposed joint null distribution and thereby corresponding multiple testing procedures depends on the user to specify the correct marginal null distribution Q_{0j} satisfying (1) for finite samples. As a consequence, for small sample sizes, exact (permutation) marginal, null distributions are preferred.

As general finite sample null distribution for a continuous T_n we propose the distribution $Q_{0n}(P)$ of $\tilde{T}_n$ defined by

$$\tilde{T}_n(j) = Q_{0j}^{-1} Q_{nj}(T_n(j)) \; j = 1, \ldots, m, \tag{2}$$

where $Q_{0j}^{-1} Q_{nj}$ is a quantile-quantile function which maps a quantile of the marginal distribution Q_{nj} of $T_n(j)$ into the corresponding quantiles of the desired marginal null distribution Q_{0j}. The j-th marginal distribution of $Q_{0n}(P)$ is exactly equal to Q_{0j}, $j = 1 \ldots, m$. We also note that this joint null distribution $Q_{0n}(P)$ does indeed satisfy the wished multivariate asymptotic domination condition (Dudoit et al. (2004)) stating that the sub-null distribution of $Q_{0n}(P)$ corresponding with the true null hypotheses asymptotically dominates the true sub-distribution of $Q_n(P)$ corresponding with these true null hypotheses. This shows that this null distribution is indeed an appropriate null distribution in re-sampling based multiple testing; we present this as a formal result below. For non-continuous T_n, we apply the generalized quantile-quantile function (Yu and van der Laan (2002))

$$\tilde{T}_n(j) = Q_{0j}^{-1} Q_{nj}^{\Delta}(T_n(j)) \; j = 1, \ldots, m,$$

where $Q_{nj}^{\Delta}(x) = \Delta Q_{nj}(x) + (1 - \Delta) Q_{nj}(x-)$, with Δ being an exogenous uniform $(0, 1)$ random variable.

Result 1 *(**Domination Properties of finite sample joint null distribution.**) Let Q_{0j} be a marginal null distribution, $j = 1, \ldots, m$, so that for $j \in \mathcal{S}_0$ and each x*

$$Q_{0j}^{-1} Q_{nj}(x) \geq x. \tag{3}$$

Let $T_n \sim Q_n(P)$, and $\tilde{T}_n \sim Q_{0n}(P)$ is defined by (2). For each $x = (x(j) : j \in \mathcal{S}_0)$ (point-wise)

$$Q_{0n,\mathcal{S}_0}(P)(x) - Q_{n,\mathcal{S}_0}(P)(x) \leq 0, \tag{4}$$

where $Q_{0n,\mathcal{S}_0}(P)$ and $Q_{n,\mathcal{S}_0}(P)$ denote the multivariate cumulative distribution functions of $(\tilde{T}_n(j) : j \in \mathcal{S}_0)$ and $(T_n(j) : j \in \mathcal{S}_0)$, respectively. In

DOI: 10.2202/1544-6115.1199

particular, we have that the number of false rejections under $Q_{0n}(P)$ stochastically dominates the true number of false rejections (i.e, under the true data generating distribution): for any c and x

$$Pr(V_n(c \mid Q_{0n}(P)) \leq x) - Pr(V_n(c \mid Q_n(P)) \leq x) \leq 0, \tag{5}$$

where $V_n(c \mid Q) = \sum_{j \in \mathcal{S}_0} I(Z_n(j) > c)$ for $Z_n \sim Q$. If the marginal domination condition (3) only holds asymptotically in the sense that (1) holds, then for all x

$$\limsup_{n \to \infty} Q_{0n,\mathcal{S}_0}(P)(x) - Q_{n,\mathcal{S}_0}(P)(x) \leq 0. \tag{6}$$

and

$$\limsup_{n \to \infty} Pr(V_n(c \mid Q_{0n}(P)) \leq x) - Pr(V_n(c \mid Q_n(P)) \leq x) \leq 0. \tag{7}$$

If (1) holds uniformly in x over any given bounded interval $[a, b]$, that is,

$$\liminf_{n \to \infty} \sup_{x \in [a,b]} \mid Q_{0j}^{-1} Q_{nj}(x) - x \mid \geq 0, \tag{8}$$

then (6) holds uniformly in x over any bounded set B in the sense that

$$\limsup_{n \to \infty} \sup_{x \in B} \mid Q_{0n,\mathcal{S}_0}(P)(x) - Q_{n,\mathcal{S}_0}(P)(x) \mid \leq 0. \tag{9}$$

Asymptotic joint null distribution with same asymptotic domination properties: Since this null distribution $Q_{0n}(P)$ depends on the true data generating distribution P, one will have to estimate it based on the data $X_1, \ldots, X_n$. We will estimate this distribution $Q_{0n}(P)$ with the regular bootstrap, which is described below. In order to establish that this bootstrap estimate indeed approximates a correct null distribution we will assume that there exists a limit joint null distribution $Q_0(P)$ so that

$$Q_{0n,\mathcal{S}_0}(P) \text{ converges weakly to the limit distribution } Q_{0,\mathcal{S}_0}(P), \tag{10}$$

where $Q_{0,\mathcal{S}_0}(P)$ denotes the sub-distribution of $Q_0(P)$ identified by the components in $\mathcal{S}_0$. The same two asymptotic domination properties (6) and (14) mentioned above apply now to this asymptotic null distribution $Q_0(P)$.

Result 2 *Assume the convergence in distribution of $Q_{0n}(P)$ to a limit distribution $Q_0(P)$ as stated in (10). The j-th marginal distribution of $Q_0(P)$ is Q_{0j}, $j = 1 \ldots, m$. If (1) holds, then for all x*

$$\limsup_{n\to\infty} Q_{0,\mathcal{S}_0}(P)(x) - Q_{n,\mathcal{S}_0}(P)(x) \leq 0, \tag{11}$$

and for all $x \in \{0, \ldots, m\}$ and any c

$$\limsup_{n\to\infty} Pr(V_n(c \mid Q_0(P)) \leq x) - Pr(V_n(c \mid Q_n(P)) \leq x) \leq 0. \tag{12}$$

In addition, if Q_{0n} is an estimate of $Q_0(P)$ and, given $(P_n : n \geq 1)$, $Q_{0n,\mathcal{S}_0}$ converges weakly to $Q_{0,\mathcal{S}_0}(P)$, then, given $(P_n : n \geq 1)$,

$$\limsup_{n\to\infty} Q_{0n,\mathcal{S}_0}(x) - Q_{n,\mathcal{S}_0}(P)(x) \leq 0, \tag{13}$$

and for all $x \in \{0, \ldots, m\}$ and all c

$$\limsup_{n\to\infty} Pr(V_n(c \mid Q_{0n}) \leq x) - Pr(V_n(c \mid Q_n(P)) \leq x) \leq 0. \tag{14}$$

The last statement implies that, if we consistently estimate the limit null distribution $Q_0(P)$, then the estimated null distribution also asymptotically dominates the true distribution of the test-statistics for the $\mathcal{S}_0$ components. We will now present a bootstrap estimate of $Q_{0n}(P)$ and view it as an estimate of the asymptotic null distribution $Q_0(P)$.

Bootstrap estimate of asymptotic null distribution $Q_0(P)$: Let $\tilde{P}_n$ be an estimate of the true data generating distribution P. If the model is nonparametric, then $\tilde{P}_n$ would be the empirical distribution P_n of $X_1, \ldots, X_n$. Alternatively, $\tilde{P}_n$ could be a model based estimate of the true data generating distribution P. Let $X_1^\#, \ldots, X_n^\#$ be an i.i.d sample from $\tilde{P}_n$, and let $P_n^\#$ be the empirical distribution of this bootstrap sample. Let $T_n^\# = T_n(P_n^\#)$ be the test-statistic vector computed from the bootstrap sample. Let $Q_{nj}^\#$ be marginal cumulative distribution of $T_n^\#(j)$, $j = 1, \ldots, m$, given P_n, which is thus known, and can be approximated by the Monte-Carlo cumulative distribution of a large sample $T_{nb}^\#(j)$, $b = 1, \ldots, B$, of B draws from $T_n^\#$.

As general bootstrap-based null distribution for T_n we propose the distribution $Q_{0n}^\#$ of $\tilde{T}_n^\#$ defined by

$$\tilde{T}_n^\#(j) = Q_{0j}^{-1} Q_{nj}^\#(T_n^\#(j)) \; j = 1, \ldots, m. \tag{15}$$

DOI: 10.2202/1544-6115.1199

This distribution can be approximated by the Monte-Carlo empirical distribution of a large sample $\tilde{T}^{\#}_{nb}$, $b = 1, \ldots, B$, based on B draws of $\tilde{T}^{\#}_{n}$. In order to exactly guarantee that the B draws from the j-th marginal distribution of $Q^{\#}_{0n}$ still exactly equals Q_{0j} we propose the generalized quantile-quantile function transformation (which also applies to discrete random variables), as proposed in Yu and van der Laan (2002). This transformation is defined as

$$\tilde{T}^{\#}_{nb}(j) = Q^{-1}_{0j} Q^{\#\Delta}_{nj,B}(T^{\#}_{nb}(j)), \text{ b=1,...,B,} \tag{16}$$

where $Q^{\#\Delta}_{nj,B}(x) \equiv \Delta Q^{\#}_{nj,B}(x-) + (1-\Delta) Q^{\#}_{nj,B}(x)$, $\Delta \sim U(0,1)$ is a random uniformly distributed random variable independent of the data, and $Q^{\#}_{nj,B}$ is the j-th marginal distribution of the Monte-Carlo approximation $Q^{\#}_{0n,B}$ of $Q^{\#}_{0n}$ based on B draws of $\tilde{T}^{\#}_{n}$. As shown in Lemma 2.3 and 2.4 Yu and van der Laan (2002), for each $B < \infty$, given P_n, $\tilde{T}^{\#}_{nb}(j) \sim Q_{0j}$, $j = 1, \ldots, m$.

To summarize, we have proposed the following procedure for establishing $Q^{\#}_{0n,B}$: 1) For $b = 1$ to B (B large) draw a bootstrap sample and compute the test-statistic vector $T^{\#}_{nb}$, which gives us an $m \times B$-matrix, 2) compute the marginal cumulative distribution functions $Q^{\#}_{nj,B} \equiv 1/B \sum_b I(T^{\#}_{nb}(j) \leq \cdot)$, $j = 1, \ldots, m$, 3) Transform the b-th column $T^{\#}_{nb}$ into $\tilde{T}^{\#}_{nb}(j) = Q^{-1}_{0j} Q^{\#\Delta}_{nj,B}(T^{\#}_{nb}(j))$, $j = 1, \ldots, m$, for each column $b = 1, \ldots, B$. The resulting $m \times B$-matrix represents now our proposed null distribution $Q^{\#}_{0n,B}$ whose marginal distributions equal Q_{0j}.

Asymptotic consistency of the bootstrap: For n converging to infinity, we will have that $\tilde{P}_n$ approximates the true data generating distribution P so that one expects that the multivariate uniform distribution $(Q^{\#}_{nj}(T^{\#}_{n}(j)) : j)$ (whose marginal distributions are exactly uniform $(0,1)$) and the multivariate uniform distribution of $(Q_{nj}(T_n(j)) : j)$ will be asymptotically identical (conditional on $(P_n : n \geq 1)$). That is, under regularity conditions, one will have that the $\mathcal{S}_0$-sub-distribution of the bootstrap distribution $Q^{\#}_{n0}$, given $(P_n : n \geq 1)$, converges weakly to the $\mathcal{S}_0$-sub-distribution of the wished limit distribution $Q_0(P)$. By the previous result above, it would then follow that our bootstrap null distribution $Q^{\#}_{0n}$ asymptotically dominates the true distribution $Q_n(P)$ for the $\mathcal{S}_0$-components, so that it is indeed appropriate to use this null distribution in re-sampling based multiple testing procedures.

2.1 Comparison with previously proposed joint null distribution (Dudoit, van der Laan, Pollard, 2004).

In this article we assumed that we know a marginal distribution Q_{0j} which dominates the true marginal distribution of $T_n(j)$ for each $j \in \mathcal{S}_0$. In Dudoit et al. (2004) it is assumed that there exist a null-mean $\mu_0(j)$ and null-variance $\sigma_0^2(j)$ so that for $j \in \mathcal{S}_0$, $\limsup_{n\to\infty} ET_n(j) \leq \mu_0(j)$ and $\limsup_{n\to\infty} \mathrm{VAR}(T_n(j)) \leq \sigma_0^2(j)$. Given the actual marginal distributions Q_{0j}, $j = 1, \ldots, m$, one could set μ_0 equal to the mean-vector of the marginal null distribution vector $(Q_{0j} : j = 1, \ldots, m)$, and it might often also be possible to choose $\sigma_0^2(j)$ equal to the variance of Q_{0j}. In Dudoit et al. (2004) we proposed as finite sample null distribution

$$\tilde{T}_n^*(j) = \min(1, \sigma_0(j)/\sqrt{\mathrm{VAR}T_n(j)})\,\{T_n(j) - ET_n(j)\} + \mu_0(j).$$

That is, the true distribution of T_n is shifted and scaled so that it has mean μ_0 and variance smaller or equal than σ_0^2. The scaling reduces (asymptotically) to multiplying with 1 for $j \in \mathcal{S}_0$, so that the only purpose of the scaling is to obtain marginal null distributions for the $j \notin \mathcal{S}_0$ which provide reasonable power. Just as in this article, in Dudoit et al. (2004) it is assumed that this finite sample null distribution converges weakly to a limit joint null distribution $Q_0^*(P)$. This null distribution is estimated with the bootstrap analogue, which involves simply null value shifting and scaling the actual bootstrap distribution of T_n. Both null distributions $Q_0(P)$ and $Q_0^*(P)$ asymptotically dominate the true distribution of the test-statistics for the sub-vector indexed by $\mathcal{S}_0$, and thereby provide the wished asymptotic control of a Type-I error rate.

Regarding the comparison of these two asymptotic null distributions $Q_0(P)$ and $Q_0^*(P)$ and their bootstrap estimates Q_{0n} and Q_{0n}^* we note the following.

- If the true marginal distributions of $T_n(j)$ converge to Q_{0j} for $j \in \mathcal{S}_0$, then the $\mathcal{S}_0$-sub-distributions of the limit distributions $Q_0(P)$ and $Q_0^*(P)$ are identical. In general, if the marginal distributions of $T_n(j)$ converge to a simple shift of Q_{0j} for $j \in \mathcal{S}_0$, then the $\mathcal{S}_0$-sub-distributions of $Q_0(P)$ and $Q_0^*(P)$ are identical.

- The marginal distributions of $\tilde{T}_n(j)$ and $\tilde{T}_n^*(j)$ and their marginal limit distributions are typically very different for $j \notin \mathcal{S}_0$, so that the $\mathcal{S}_0^c$-sub-distributions of $Q_0(P)$ and $Q_0^*(P)$ are very different. In particular,

DOI: 10.2202/1544-6115.1199

while the j-th marginal distribution of $Q_0(P)$ is Q_{0j}, the null-value shifted and scaled j-th marginal distributions of $Q_0^*(P)$ is not necessarily equal to Q_{0j}, $j \notin \mathcal{S}_0$.

- The bootstrap estimate Q_{0n} can be expected to be a more efficient estimate of $Q_0(P)$ than Q_{0n}^* is of $Q_0^*(P)$. To be concrete, consider the case that the $\mathcal{S}_0$-sub-distributions of $Q_0(P)$ and $Q_0^*(P)$ are identical, and compare the estimates for this common sub-distribution. The bootstrap estimate $Q_{0n,\mathcal{S}_0}$ can be viewed as an estimate of $Q_{0,\mathcal{S}_0}(P)$ in the model in which all marginal distributions Q_{0j} are given, while the bootstrap estimate Q_{0n}^* ignores this knowledge about the marginal distributions, and indeed its finite sample j-th marginal distribution is different from Q_{0j}. For example, if $Q_0(P)$ is multivariate normal, then the marginal distributions of Q_{0n} are fixed and equal to the marginal distributions of $Q_0(P)$, while the marginal distributions of Q_{0n}^* are subject to finite sample variability.

- The fact that the marginal distributions of Q_{0n} are exactly identical to the user supplied Q_{0j} is particularly appealing if these marginal distributions are actually known. For example, if the null hypothesis H_{0j} states that X_j is independent of an outcome Y, then one can set $Q_{0j} = Q_{0j,n}$ equal to the permutation distribution of the test-statistic $T_n(j)$. It is known that, if the null hypothesis is true, then this permutation distribution of $T_n(j)$ is actually the exact true conditional distribution of $T_n(j)$, given the marginal empirical distribution of X_j and the marginal empirical distribution of Y. Thus, in this case, the marginal distributions of Q_{0n} are exact (a marginal test would provide an exact p-value). Alternatively, if one uses a t-statistic for a test $H_{0j} : \mu(j) \leq \mu^0(j)$, then the marginal limit distribution Q_{0j} of Q_0 is known to be $N(0,1)$ (by the central limit theorem).

2.2 Comparing re-sampling based multiple testing with marginal p-value multiple testing.

Because of the fact that our re-sampling based multiple testing procedure can be based on a joint null distributions with user supplied marginal distributions, we can now provide a valid comparison between a multiple testing procedure based on marginal null distributions (i.e., p-values) and our cor-

responding re-sampling based multiple testing procedure using our joint null distribution with the same marginal null distributions. Consider such a multiple testing method based on marginal p-values based on knowing for each null hypothesis a dominating marginal distribution of the test-statistic under the null hypothesis. In addition, assume that this marginal-p-value multiple testing method is not based on additional assumptions so that it is a method which controls the wished type-I error rate under any kind of joint distribution between the test-statistics. For example, this might be the multiple testing procedure which rejects any null hypothesis for which the marginal p-value is smaller or equal than α/m (i.e., the Bonferoni procedure), or the corresponding so called Holmes step down method, where the p-values are calculated under a known marginal null distribution (e.g., $N(0,1)$). Both of these procedures are known to provide asymptotic control of the family wise error under any data generating distribution. The re-sampling based analogues of the Bonferoni and Holmes methods would now be the single step method based on controlling the FWE under our joint null distribution with the same marginal null distributions, and the step down method based on this same joint null distribution, respectively. Since the Bonferoni cut-offs are valid under any joint distribution and the re-sampling based multiple testing methods are based on cut-offs that are valid under an estimate of the true joint distribution of the test-statistics corresponding with the true nulls, it follows that the single step and step-down re-sampling based multiple testing methods are less conservative than the Bonferoni procedure and Holmes step down method, respectively.

Of course, the same argument applies in general. We remind the reader that, given a multiple testing procedure indexed by a nominal level of a Type-I error rate, an adjusted-p-value for a particular test-statistic value $t_n(j)$ for $T_n(j)$ is the smallest α at which the multiple testing procedure rejects H_{0j}, $j = 1, \ldots, m$. As a consequence of the above argument, an adjusted p-value for a re-sampling-based multiple testing procedure based on our joint null distribution is smaller than the adjusted-p-value for the corresponding multiple testing procedure based on the marginal null distributions only.

DOI: 10.2202/1544-6115.1199

3 Formal theoretical framework to establish asymptotic control of Type-I error.

The fundamental theorems 1-4 in Dudoit et al. (2004) are concerned with establishing asymptotic control of a type-I error rate for a single step multiple testing procedure controlling this type-I error rate under a null distribution $Q_0(P)$ or an estimate thereof. These theorems rely solely on the asymptotic domination condition of the limit null distribution $Q_0(P)$ (Theorem 1 and 2) relative to the true distribution of the test-statistics $Q_n(P)$, and on weak convergence of an estimated null distribution Q_{0n} (such as our bootstrap distribution $Q_{0n}^{\#}$) to $Q_0(P)$ (Theorem 3 and 4). Therefore a simple application of these theorems provides us with the wished results for single step re-sampling based multiple testing procedures. In order to provide the reader with a concrete presentation of a result in this article we will here just state the precise statement for our null distribution $Q_0(P)$ analogue to Theorem 1 of Dudoit et al. (2004). This theorem is an immediate corollary of Theorem 1 in Dudoit et al. (2004) and the asymptotic domination condition (1) for our joint null distribution Q_0.

Theorem 1 *Assume the convergence in distribution of $Q_{0n}(P)$ to a limit distribution $Q_0 = Q_0(P)$ as stated in (10), and assume condition (1).*

Consider a type-I error rate $\theta(F_{V_n}) \in \mathbb{R}$, F_{V_n} being the cumulative distribution function of the number of false rejections, V_n, of a given multiple testing procedure. Assume that this real valued parameter $F \to \theta(F)$ on cumulative distribution functions on $\{1, \ldots, m\}$ satisfies the monotonicity assumption AMI and the continuity assumption ACI as defined in Dudoit et al. (2004): e.g., $\theta(F) = 1 - F(0)$ (FWE). Let $(d_j(Q_0, \alpha) : j = 1, \ldots, m)$ with $d_j(Q_0, \delta) = \inf\{z : Q_{0j}(z) \geq \delta\}$, $j = 1, \ldots, m$, be the common-quantile cut-off rule. Let $c(Q_0, \alpha) = d(Q_0, \delta_0(\alpha))$ be the cut-off vector which yields the wished control under Q_0, which is defined by

$$\delta_0(\alpha) = \inf\{\delta : \theta(F_{R_0(d(Q_0,\delta))}) \leq \alpha\},$$

and $R_0(d) \equiv \sum_{j=1}^{m} I(Z_j > d_j)$ is the number of rejections at cut-off vector d with $Z \sim Q_0$. Consider the single step multiple testing procedure

$$\mathcal{S}(T_n, Q_0, \alpha) \equiv \{j : T_n(j) > d_j(Q_0, \delta_0(\alpha))\}.$$

This multiple testing procedure $\mathcal{S}(T_n, Q_0, \alpha)$ *provides asymptotic control of the Type-I error rate* $\theta(F_{V_n})$ *at level* α*. That is,*

$$\limsup_{n\to\infty} \theta(F_{V_n}) \leq \alpha,$$

where $V_n = \sum_{j=1}^{n} I(T_n(j) > d_j(Q_0, \delta_0(\alpha)), j \in \mathcal{S}_0)$ *denotes the number of false positives, and* $T_n \sim Q_n(P)$ *follows the true distribution* $Q_n(P)$*.*

Similarly, applications of Theorem 3 and 4 in Dudoit et al. (2004) imply that, if the bootstrap estimated distribution $Q_{0n}^{\#}$, given $(P_n : n \geq 1)$, weakly converges to $Q_0 = Q_0(P)$, then the multiple testing procedure $\mathcal{S}(T_n, Q_{0n}^{\#}, \alpha)$ also provides asymptotic control of $\theta(F_{V_n})$ at level α, under some mild continuity conditions on $Q_0(P)$.

As a consequence, our null distribution $Q_{0n}^{\#}$ provides asymptotic control of the type-I error rate for single step multiple testing procedures, as in Dudoit et al. (2004). In addition, an application of the theorems in van der Laan et al. (2004c) shows that step-down methods for control of FWE, as presented in van der Laan et al. (2004c), based on $Q_{0n}^{\#}$, asymptotically control the FWE. An application of the theorems in van der Laan et al. (2004b) demonstrates that the augmentation methods for controlling GFWER and TPPFP(q) based on $Q_{0n}^{\#}$ are asymptotically valid. Finally, the empirical Bayes re-sampling based multiple testing procedure controlling TPPFP(q) based on $Q_{0n}^{\#}$ provides asymptotic control of the TPPFP(q) (van der Laan et al. (2004a)).

Beyond the asymptotic validity of any re-sampling based multiple testing method based on the joint null distribution $Q_{0n}^{\#}$, we suggest that this new null distribution will provide significant *finite sample* improvements to all these re-sampling based multiple testing procedures due to the fact that we can completely control the marginal null distributions. In particular, if a known marginal null distribution of the test-statistic is known to provide exact marginal Type-I error control (e.g., permutation distribution for testing independence between two random variables, or the students distributions for a t-statistic), then our joint null distribution can be expected to provide more accurate control in finite samples than the previously proposed joint null distribution.

Remark. Regarding the results for Type-I error control such as the one stated in Theorem 1, since the results only concern the distribution of the $\mathcal{S}_0$-sub-vector of the test statistics, one only needs that the $\mathcal{S}_0$-sub-distribution

DOI: 10.2202/1544-6115.1199

of Q_{0n} (or $Q_{0n}^{\#}$) converges to the $\mathcal{S}_0$-sub-distribution of $Q_0(P)$. In certain applications, the latter is easier to show than the weak convergence to a complete m-dimensional limit distribution.

4 Two general examples: t-statistics and Chi-square statistics.

In this section, we present the new MTP for two general classes of multiple testing problems. First, we consider null hypotheses stating that a real valued parameter is smaller or equal than a hypothesized value, where we allow any kind of real valued parameter which can be estimated at a root-n-rate. For example, one might wish to test the null hypothesis that cancer cells have a particular gene expression that is less than or equal to non-cancer cells. In the second subsection we consider null hypotheses stating that K real valued parameters are equal. For example, one might wish to test the null hypothesis that 3 different sub-types of cancer have equal mean gene expression.

4.1 Testing real valued parameters.

Let $\Psi(P)(j)$ be a real valued path-wise differentiable parameter of P, and consider the null hypothesis $H_{0j} : \Psi(P)(j) \leq \psi_{0j}$ for a null-value ψ_{0j}, $j = 1, \ldots, m$. Given an asymptotically linear estimator $\psi_n(j)$ of $\Psi(P)(j)$, let $T_n(j) = (\psi_n(j) - \psi_{0j})/\sigma_n(j)$ be the test-statistic for testing the null hypothesis H_{0j}, where $\sigma_n(j)$ is the standard error of $\psi_n(j)$. Let $IC_j(O \mid P)$ be the influence curve of the estimator $\psi_n(j)$ at P, $j = 1, \ldots, m$.

Since $T_n(j) = (\psi_n(j) - \Psi(P)(j))/\sigma_n(j) + (\Psi(P)(j) - \psi_{0j})/\sigma_n(j)$ and the first term converges to a $N(0,1)$ by the central limit theorem, it follows that, if $j \in \mathcal{S}_0$, then $T_n(j)$ is asymptotically dominated by a standard normal (i.e., $N(0,1)$). Therefore, we should set $Q_{0j} = \Phi$ equal to the cumulative distribution function Φ of a standard normally distributed random variable. It also follows that the true marginal distribution Q_{nj} of $T_n(j)$ can be approximated by a normal distribution with mean $(\Psi(P)(j) - \psi_{0j})/\sigma_n(j)$ and variance 1. As a consequence, it follows that $Q_{0j}^{-1} Q_{nj}(T_n(j)) \approx (\psi_n(j) - \Psi(P)(j))/\sigma_n(j)$ corresponds in first order with subtracting from $T_n(j)$ its shift $(\Psi(P)(j) - \psi_{0j})/\sigma_n(j)$. However, the quantile-quantile function transformation actually guarantees that $Q_{0j}^{-1} Q_{nj}(T_n(j)) = \Phi^{-1} Q_{nj}(T_n(j)) \sim N(0,1)$ exactly. Our

bootstrap joint null distribution, $Q_{0n}^{\#}$, is the conditional distribution of $\tilde{T}_n^{\#} \equiv (\Phi^{-1}Q_{nj}^{\#}(T_n^{\#}(j)), j = 1, \ldots, m)$, given P_n, where $T_n^{\#} = (\psi_n^{\#} - \psi_0)/\sigma_n^{\#}$ (using vector notation) is the test-statistic vector based on a bootstrap sample $O_1^{\#}, \ldots, O_n^{\#} \sim P_n$, and $Q_{nj}^{\#}$ is the marginal distribution of $T_n^{\#}(j)$. In the next result we prove that, under mild regularity conditions, the conditional distribution of $\tilde{T}_n^{\#}$, given $(P_n : n \geq 1)$, converges weakly to $Q_0(P) = N(0, \Sigma(P))$, where $\Sigma(P)$ is the correlation matrix of the vector influence curve $IC(O \mid P)$. This proves that our joint null distribution is asymptotically equivalent with the joint null distribution proposed in Pollard and van der Laan (2003) and the null-value shifted joint null distribution in Dudoit et al. (2004). In particular, this shows that any of the theorems establishing Type-I error control in these latter papers also apply to this bootstrap null distribution $Q_{0n}^{\#}$.

Theorem 2 *Assume that ψ_n is an asymptotically linear estimator of $\Psi(P) = (\Psi_1(P), \ldots, \Psi_m(P))$ with influence curve $IC(O \mid P)$. Let $\Sigma(P)$ be the correlation matrix of $IC(O \mid P)$, and let σ_n be an estimator of the standard error of ψ_n so that $Z_n \equiv (\psi_n - \psi)/\sigma_n$ converges weakly to $Q_0(P) = N(0, \Sigma(P))$. Assume that $\sigma_n^{\#} - \sigma_n$ converges to zero for $n \to \infty$ a.s., and that, given $(P_n : n \geq 1)$, the distribution of $(\psi_n^{\#} - \psi_n)/\sigma_n^{\#}$ converges to the same limit distribution $Q_0(P) = N(0, \Sigma(P))$ as $(\psi_n - \psi)/\sigma_n$. Then, given $(P_n : n \geq 1)$,*

$$(\tilde{T}_n^{\#}(j) : j) = (\Phi^{-1}Q_{nj}^{\#}(T_n^{\#}(j)) : j) \stackrel{D}{\Rightarrow} Q_0(P).$$

Proof: Let $Z_n^{\#} = (\psi_n^{\#} - \psi_n)/\sigma_n^{\#}$, and let $d_n = (\psi_n - \psi_0)/\sigma_n^{\#}$. We have, conditional on P_n,

$$Q_{nj}^{\#}(x) = F_{Z_n^{\#}(j)}(x - d_n(j)) + o(1),$$

where F_Z denotes the cumulative distribution function of Z, conditional on $(P_n : n \geq 1)$. The $o(1)$ term converges to zero uniformly in x because $\sigma_n^{\#}$ behaves as a non-random σ_n in the sense that $\sigma_n^{\#} - \sigma_n \to 0$ for $n \to \infty$, a.s. Because $Z_n^{\#}$, given $(P_n : n \geq 1)$, converges weakly to $N(0, \Sigma(P))$, and point-wise convergence of monotone functions to a continuous limit implies uniform convergence, we have that for $n \to \infty$

$$\sup_x \mid F_{Z_n^{\#}(j)}(x) - \Phi(x) \mid \to 0.$$

Thus, for $n \to \infty$,

$$\sup_x \mid Q_{nj}^{\#}(x) - \Phi(x - d_n(j)) \mid \to 0.$$

DOI: 10.2202/1544-6115.1199

As a consequence, it follows that for $n \to \infty$ and any bounded interval $[a, b]$

$$\sup_{x \in [a,b]} \mid \Phi^{-1} Q_{nj}^{\#}(x) - (x - d_n(j)) \mid \to 0.$$

Thus

$$\tilde{T}_n^{\#}(j) = T_n^{\#}(j) - d_n(j) + o_P(1) = Z_n^{\#}(j) + o_p(1).$$

This shows that, conditional on $(P_n : n \geq 1)$, $\tilde{T}_n^{\#}$ converges to the same limit distribution, $Q_0(P) = N(0, \Sigma(P))$, as $Z_n^{\#}$. This completes the proof. □

4.2 Testing equality of real valued parameters.

We observe n i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P$. Let $\Psi_{jk}(P)$, $j = 1, \ldots, m$, $k = 1, \ldots, K$, be a collection of real valued path-wise differentiable parameters. Suppose that the null hypotheses of interest are $H_0(j) : \Psi_{j1}(P) = \Psi_{j2}(P) \ldots = \Psi_{jK}(P)$, $j = 1, \ldots, m$. That is, for each j, we wish to test equality of the K corresponding real valued parameters $\Psi_{j1}(P), \ldots, \Psi_{jK}(P)$. For example, consider the case that $O = (Y, L)$, where $Y = (Y(1), \ldots, Y(m))$ is an outcome vector, $L \in \{1, \ldots, K\}$ is a group label, and we are concerned with testing equality of means across the K groups: $H_{0j} : E(Y(j) \mid L = 1) = \ldots = E(Y(j) \mid L = K)$, $j = 1, \ldots, m$. In this example, $\Psi_{jk}(P) = E_P(Y(j) \mid L = k)$, $k = 1, \ldots, K$, $j = 1, \ldots, m$.

Let $\psi_{jk,n}$ be an asymptotically linear estimator of $\Psi_{jk}(P)$ with influence curve $IC_{jk}(O \mid P)$. Let $\bar{\psi}_{jn}$ be an asymptotically linear estimator of $\bar{\Psi}_j(P) = \frac{1}{K} \sum_{k=1}^{K} \Psi_{jk}(P)$ with influence curve $\bar{IC}_j(O \mid P)$, $j = 1, \ldots, m$. For example, $\bar{\psi}_{jn} = \frac{1}{K} \sum_{k=1}^{K} \psi_{jk,n}$, or one might wish to use weights inversely proportional to a variance estimator of $\psi_{jk,n}$, and use a corresponding weighted average. Now, note that, under H_{0j}, $(\psi_{jk,n} - \bar{\psi}_{jn} : k = 1, \ldots, K)$ is an asymptotically linear estimator of $(\Psi_{jk}(P) - \bar{\Psi}_j(P) : k = 1, \ldots, K)$ with influence curve $IC_j(O \mid P) \equiv (IC_{jk}(O \mid P) - \bar{IC}_j(O \mid P) : k = 1, \ldots, K)$. Let $\Sigma_j(P)$ be the $K \times K$ covariance matrix of this K-dimensional influence curve $IC_j(O \mid P)$, and let Σ_{jn} be an estimate of this covariance matrix. By the central limit theorem we have, under H_{0j},

$$\Sigma_{jn}^{-0.5}(\psi_{jk,n} - \bar{\psi}_{jn} : k = 1, \ldots, K-1) \stackrel{D}{\Rightarrow} N(0, I_{K-1 \times K-1}).$$

Therefore we propose as test-statistic the squared Euclidean norm of the latter quantity:

$$T_n(j) \equiv \| \Sigma_{jn}^{-0.5}(\psi_{jk,n} - \bar{\psi}_{jn} : k = 1, \ldots, K-1) \|^2 .$$

If H_{0j} is true, then the marginal distribution of $T_n(j)$ converges to a $\mathcal{X}^2$-distribution with $K-1$ degrees of freedom. Let G be the cumulative distribution function of this $\mathcal{X}^2$-distribution.

Let $O_1^\#, \ldots, O_n^\#$ be i.i.d. draws from the empirical distribution function P_n of $O_1, \ldots, O_n$ or a model based estimate $\tilde{P}_n$ of P. Let Q_{nj} be the true marginal distribution of $T_n(j)$, $Q_{nj}^\#$ be the bootstrap marginal distribution of $T_n^\#(j)$ based on $O_1^\#, \ldots, O_n^\#$, given P_n. Our proposed joint null distribution, Q_{0n}, for the test statistic T_n is now the distribution of

$$\tilde{T}_n = (\tilde{T}_n(j) : j) = (G^{-1}Q_{nj}(T_n(j)) : j),$$

and its bootstrap estimate, $Q_{0n}^\#$, is the distribution of

$$\tilde{T}_n^\# = (\tilde{T}_n^\#(j) : j) = (G^{-1}Q_{nj}^\#(T_n^\#(j)) : j), \text{ given } P_n.$$

Note that the marginal distributions of Q_{0n} and $Q_{0n}^\#$ are all exactly equal to the Chi-square distribution G with $K-1$ degrees of freedom.

We will now establish the limit distribution of $(\tilde{T}_n(j) : j \in \mathcal{S}_0)$ and show that the proposed bootstrap distribution $(\tilde{T}_n^\#(j) : j \in \mathcal{S}_0)$ converges weakly to this limit distribution. Application of the previous stated and mentioned theorems now show that re-sampling based multiple testing procedures based on the bootstrap null distribution $Q_{0n}^\#$ asymptotically control the wished Type-I error rate.

Theorem 3 *We observe n i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P$. Let $\Psi_{jk}(P)$, $j = 1, \ldots, m$, $k = 1, \ldots, K$, be a collection of real valued path-wise differentiable parameters. Suppose that the null hypotheses of interest are $H_0(j) : \Psi_{j1}(P) = \Psi_{j2}(P) \ldots = \Psi_{jK}(P)$, $j = 1, \ldots, m$.*

Let $\psi_{jk,n}$ be an asymptotically linear estimator of $\Psi_{jk}(P)$ with influence curve $IC_{jk}(O \mid P)$. Let $\bar{\psi}_{jn}$ be an asymptotically linear estimator of $\bar{\Psi}_j(P) = \frac{1}{K}\sum_{k=1}^K \Psi_{jk}(P)$ with influence curve $\bar{IC}_j(O \mid P)$, $j = 1, \ldots, m$. Let $IC_j(O \mid P) \equiv (IC_{jk}(O \mid P) - \bar{IC}_j(O \mid P) : k = 1, \ldots, K-1)$. Let $\Sigma_j(P)$ be the $K-1 \times K-1$ covariance matrix of this $K-1$-dimensional influence curve $IC_j(O \mid P)$, assume it is invertible, and let Σ_{jn} be a consistent estimate of this covariance matrix. Define as test-statistics

$$T_n(j) \equiv \| \Sigma_{jn}^{-0.5}(\psi_{jk,n} - \bar{\psi}_{jn} : k = 1, \ldots, K-1) \|^2 \quad j = 1, \ldots, m.$$

Let G be the cumulative distribution function of a $\mathcal{X}^2$-distribution with $K-1$ degrees of freedom.

DOI: 10.2202/1544-6115.1199

Let $IC^*_{jk}(O \mid P) \equiv IC_{jk}(O \mid P) - \bar{IC}_j(O \mid P)$ *denote the influence curve of* $\psi_{jk,n} - \bar{\psi}_{jn}$ *as an estimator of* $\Psi_{jk}(P) - \bar{\Psi}_j(P)$, $j = 1, \ldots, m$, $k = 1, \ldots, K-1$. *Let* $IC^*(O \mid P) \equiv (IC^*_{jk}(O \mid P) : j, k)$ *be the corresponding* $m \times K - 1$*-vector influence curve. Let* $(Z(j,k) : j,k) \sim N(0, \Sigma^*(P))$, *where* $\Sigma^*(P) \equiv E(IC^*(O \mid P) IC^{*\top}(O \mid P))$, *and its elements are denoted by* $\Sigma^*(P)((j_1,k_1),(j_2,k_2)) = E_P IC^*_{j_1k_1}(O \mid P) IC^*_{j_2,k_2}(O \mid P)$. *Let* $Q_0(P)$ *be the distribution of the random variable* $(Z(j) : j)$ *defined as*

$$Z(j) = \| \Sigma_j^{-0.5}(Z(j,k) : k = 1, \ldots, K-1) \|^2, \ j = 1, \ldots, m.$$

Consider

$$\tilde{T}_n = (\tilde{T}_n(j) : j) = (G^{-1} Q_{nj}(T_n(j)) : j).$$

Then, $(\tilde{T}_n(j) : j \in \mathcal{S}_0) \stackrel{D}{\Rightarrow} (Z(j) : j \in \mathcal{S}_0)$.

In addition, assume that $\Sigma^{\#}_{jn}$ *is a consistent bootstrap estimator of* Σ_j, *and the conditional distribution of* $\sqrt{n}(\psi^{\#}_{jk,n} - \bar{\psi}^{\#}_{jn} : j,k)$, *given* $(P_n : n \geq 1)$, *converges to* $N(0, \Sigma^*(P))$. *Then* $(\tilde{T}^{\#}_n(j) : j \in \mathcal{S}_0)$ *converges weakly to* $(Z(j) : j \in \mathcal{S}_0)$, *given* $(P_n : n \geq 1)$.

Note that the latter condition just states that the bootstrap is asymptotically consistent in estimating the limit distribution of $\sqrt{n}(\psi_{jk,n} - \bar{\psi}_{jn} : j,k)$, which is therefore a mild regularity condition (see e.g., Gill (1989)).

Proof. For $j \in \mathcal{S}_0$, it follows that Q_{nj} converges uniformly to G, and thereby that for any bounded interval $[a,b]$ $\sup_{x \in [a,b]} \mid G^{-1} Q_{nj}(x) - x \mid \rightarrow 0$ for $n \rightarrow \infty$. Thus, $(\tilde{T}_n(j) : j \in \mathcal{S}_0) = (T_n(j) : j \in \mathcal{S}_0) + o_P(1)$. By the continuous mapping theorem (van der Vaart and Wellner (1996)) and the weak convergence of $\sqrt{n}(\psi_{jk,n} - \bar{\psi}_{jn} : j,k)$ to a multivariate normal distribution $N(0, \Sigma^*(P))$, it follows that $(T_n(j) : j \in \mathcal{S}_0)$ converges weakly to $(Z(j) : j \in \mathcal{S}_0)$, where the latter distribution is specified in the theorem. Similarly, by the continuous mapping theorem and by the weak convergence of the bootstrap distribution $\sqrt{n}(\psi^{\#}_{jk,n} - \bar{\psi}^{\#}_{jn} : j,k)$ to the same multivariate normal distribution $N(0, \Sigma^*(P))$, it follows that, conditional on $(P_n : n \geq 1)$, $(\tilde{T}^{\#}_n(j) : j \in \mathcal{S}_0)$ converges weakly to $(Z(j) : j \in \mathcal{S}_0)$. Thus, the $\mathcal{S}_0$-sub-distribution of $\tilde{T}_n$ and $\tilde{T}^{\#}_n$ (conditional on $(P_n : n \geq 1)$) converge to the $\mathcal{S}_0$ sub-distribution of $Q_0(P)$ specified above. □

5 A simulation study.

Simulated data was used to examine the relative finite sample performance of the newly proposed quantile transformation method relative to the null-centered, re-scaled bootstrap MTP as well as that based on marginal p-values (e.g., the Bonferroni method). The data generating mechanism was intended to provide situations for which Pearson χ^2 test is a valid test-statistic for testing the independence of two variables, Z and Y. Let $X = (Z, Y)$. In addition, we also wanted to engender dependence among test statistics to examine whether the quantile-method gained power over MTP's based strictly on marginal p-values (such as Bonferroni) that are only sharp under independence. The data-generating mechanism is defined as follows:

- $X_i = (Z_i, Y_{i1}, \ldots, Y_{i100})$, where Z_i was uniform over (0,1,2), $i = 1, \ldots, n$, $n = 99$.
- Y_{ij} was binary, where $P(Y_{ij} = 1 \mid Z_i) = 1/(1 + exp(-(\beta_{0i} + \beta_{1j} Z_i)))$.
- $\beta_{0i} \sim N(-0.57, 10)$ which means that every subject has a random intercept. This results in high positive correlation of the Y_{ij}'s measured on the same subject, i, which engenders correlation of the test statistics, $T_n(j), j = 1, \ldots, 100$.
- For $j = 1, \ldots, 75$, let $\beta_{1j} = 0$. For the remaining j it is fixed at a constant value, 0.40. The null hypotheses of interest are H_{0j} : Y_j independent of Z, $j = 1, \ldots, 100$. Note that the first 75 null hypotheses are true and the last 25 are false.
- The $T_n(j)$ are classic Pearson χ^2 statistics testing independence, which, under the null hypothesis, are asymptotically χ^2 distributed with 2 degrees of freedom (a binary variable, Y, vs. a variable with 3 categories, Z).

To summarize, we generated data that resulted in 100 dependent test statistics, for which we know the null hypothesis is true for 75 of them. In addition, we also know that if the null hypothesis is true, then the actual marginal distribution of the test statistics should be χ^2 distributed with 2 degrees of freedom. Thus we can choose the correct Q_{0j}. For each simulated data set we used the bootstrap to derive the null-centered, re-scaled test statistic and used a single-step approach to control family wise error rate

DOI: 10.2202/1544-6115.1199

(FWER) defined as the probability of rejecting any true null hypothesis. This method for controlling FWER using a re-sampling based MTP is described in Dudoit et al. (2004), which involves randomly re-sampling the independent units with replacement, calculating the Pearson χ^2 test statistic for each of the 100 tests, and repeating this B times ($B = 5000$ in this case). This results in a matrix of 100 rows and B columns. After re-scaling them to have the correct variance (a maximum of 2 times the degrees of freedom, or 4) and centering them to have mean equal to the degrees of freedom (2), one obtains the Monte-Carlo approximation of the proposed null distribution in Dudoit et al. (2004) based on 5000 replicates. We are now concerned with finding a common cut-off for the test-statistics so that under this null distribution the FWER is equal to 0.05. Thus, one finds the maximum for each column - the maximum over each of the 100 null-shifted, re-scaled test statistics, and one selects the 0.95 quantile of the obtained 5000 maxima. To control the FWER at 0.05, we reject the null hypothesis only if the observed $T_n(j)$ is greater than the 0.95 quantile of these maximums. The quantile method works the same, but the centering and scaling is now replaced by the quantile transformation on each row of orignal (before null-centering and re-scaling) matrix of bootstrapped test statistics, where Q_{0j}^{-1} is now the inverse of the CDF of the χ^2, df=2 distribution.

Finally, the Bonferroni method is equivalent to rejecting the null hypothesis if the p-value is less than the desired FWER of 0.05 divided by the number of tests: in this case reject the null if $1 - F_{\chi^2,df=2}(T_n(j)) < 0.0005$. For each simulated data set, the number of falsely rejected and accepted null hypotheses is recorded and the probability of making such mistakes estimated from repeated simulations (1000 in this case) provides both an estimate of the Type I error control and the power of these competing procedures.

First, for a single simulation, we compared the density of the null-centered re-scaled test statistics, the quantile transformed test statistics and the desired marginal null distribution (χ^2) among those test statistics for which the null is true. Figure 1 presents the smoothed (kernel) density estimators, both over the range of $T_n(j)$ and then for just the right tail. As one can see, the null-centered, re-scaled distribution does a poor job of approximating the χ^2 distribution, whereas the quantile method is perfect (curves overlap and can not be distinguished). For the repeated simulations, when the desired FWER was set at 0.05, the results show that the FWER based on the null-centered, re-scaled MTP is anti-conservative ($FWER > 0.05$), whereas the Bonferroni method is very conservative ($FWER = 0.005$). However,

the FWER based upon quantile transformed null distribution is reasonably sharp (FWER = 0.04). A single step method, whatever null distribution is used, should be conservative since not all null hypotheses are true. We also note that the quantile-transformation method has 10 times the power of the Bonferroni method. Thus, the simulations seem to confirm what the above discussion predicts - that a MTP based on a quantile-quantile function transformed bootstrap distribution of the test-statistics can remedy practical problems with the original re-sampling based procedures, in that good error control can be achieved with a practical number bootstrap repeats. We note that as B gets larger, by theory, the FWER for the null-value centering and scaled null distribution will also achieve the wished asymptotic control of the family wise error, but B must be very large to get accurate control in the extreme tails, and also the sample size will probably need to be larger. The quantile transformed bootstrap null distribution appears to provide a fix to these practical problems with the re-sampling based methodology. The bottom-line is a new technique that gives the optimal marginal inference and provides powerful and generally valid MTP's by taking advantage of the dependence among test statistics.

6 A data example.

We examine the same competing MTP's on a data set of the association of single nucleotide poly-morphisms (SNPs) in the ghrelin (GHRL) and neuropeptide Y (NPY) genes and a form of cancer, non-Hodgkin lymphoma (NHL). There is some biologic plausibility that poly-morphisms in these genes might be related to NHL (see Skibola et al. (2005) for more detail). We examined the association of 11 SNP's, each of which have 3 levels (homozygous wild-type, heterozygous, and homozygous mutant) and NHL. For each of these SNP's we also examine the association with 4 different sub-types of NHL: 1) all NHL cases, 2) diffuse large-cell lymphoma (DLCL), 3) follicular lymphoma (FL) and 4) cancers that are neither DLCL nor FL, resulting in a total of 44 tests. The samples size is $n = 992$. As in the simulations, the null hypotheses of interest are the independence of the SNP and NHL across the SNPs, and for each null hypothesis we use the Pearson χ^2 test. Table 1 has the raw and adjusted p-values based on the same three procedures performed in the simulations. As one can see, though no comparisons would be significant at an FWER of 0.05, the quantile transformation method appears to

DOI: 10.2202/1544-6115.1199

Figure 1: Density of null distributions: null-centered, rescaled bootstrap, quantile-transformed and the theoretical. A is over entire range, B is the right tail.

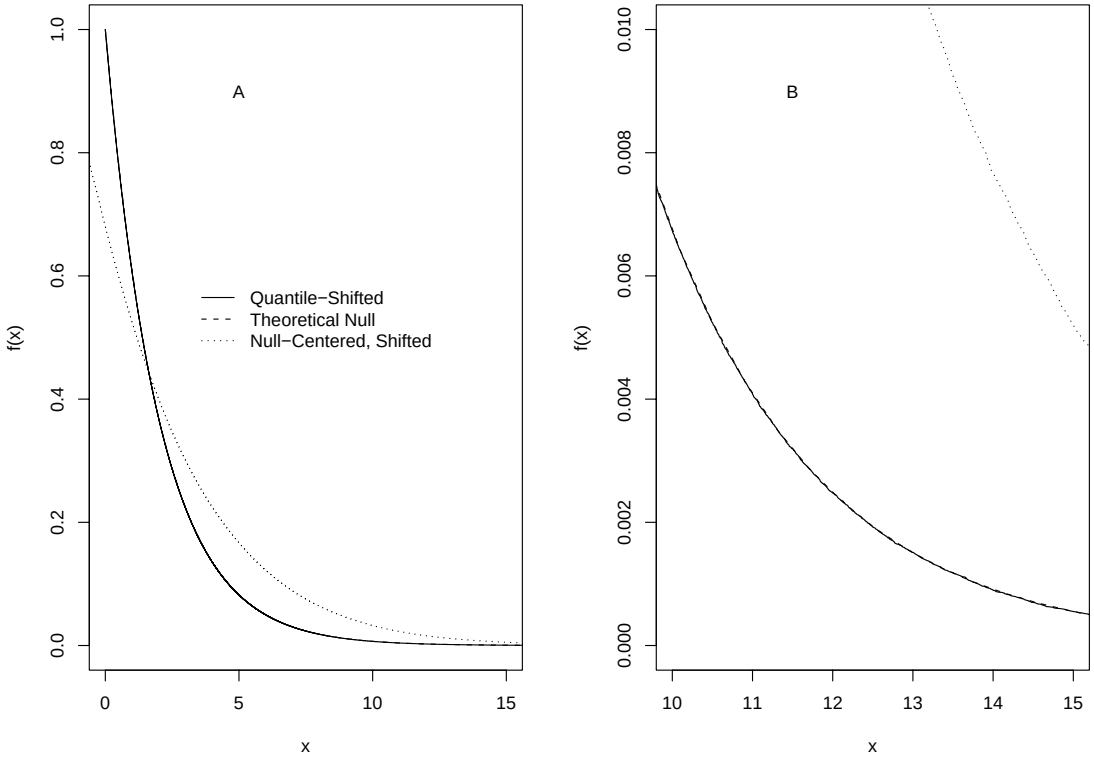

Table 1: An ordered list (by p-value) of the top 10 SNP's/NHL sub-group with adjusted p-values using 3 methods: raw, Bonferroni (B), Null-centered re-scaled bootstrap (NCRB) and quantile transformation (QT)

SNP-NHL type	χ^2 statistic	raw	B	NCRB	QT
py5671ct-all	9.59	0.0083	0.365	0.306	0.239
ghrl4427ag-DBLCL	8.64	0.0133	0.585	0.454	0.353
py5671ct-FCC	8.33	0.0155	0.682	0.503	0.396
py485ct-FCC	7.27	0.0264	1	0.717	0.566
npy1258ga-all	6.41	0.0406	1	0.865	0.711
eptina19g-all	6.23	0.0444	1	0.891	0.740
py1258ga-FCC	6.10	0.0473	1	0.907	0.761
py485ct-all	6.01	0.0496	1	0.917	0.779
npy1128tc-FCC	5.23	0.0731	1	0.980	0.889
leptina19g-FCC	4.59	0.1009	1	0.997	0.951

be less conservative than the Bonferroni method and also the null-centered, re-scaled bootstrap approach.

7 Discussion.

In most marginal testing problems the wished (e.g., most powerful) marginal null distribution is known. Therefore, methods based on marginal p-values have been attractive because of that very reason, but users are also concerned about the fact that these methods are only sharp under complete independence. The re-sampling based multiple testing methodology based on a data generating null distribution in Westfall and Young (1993) controlled the choice of marginal null distributions, and aims to estimate a valid joint distribution for the test-statistics, but it relies on a very restrictive subset pivotality condition. Therefore it can only be applied to a limited set of multiple testing problems. The recently proposed re-sampling-based multiple testing methods based on the null-value centered and scaled bootstrap distribution of the test-statistics resolves the restrictive subset-pivotality condition, but its finite sample performance can suffer due to the fact that the marginal null distributions cannot be controlled. As a consequence, this method, though asymptotically always valid, might require larger sample sizes and large number of bootstrap replicates. We view our newly proposed resampling based

DOI: 10.2202/1544-6115.1199

multiple testing methodology based on the quantile-tranformed null distribution as the method solving all the above concerns: it provides asymptotically sharp control, it does not rely on the subset pivotality condition, and it still controls the choice of marginal null distributions. As a consequence, it will typically outperform the other methods in finite samples and asymptotically.

APPENDIX - Glossary of Notation

$X_1, \ldots, X_n$ the data on n i.i.d units having distribution, $X \sim P$.

P_n is the empirical distribution of X.

$H_{0j}, j = 1, \ldots, m$ are the null hypotheses of interest.

$T_n = (T_n(1), \ldots, T_n(m))$ vector of observed test-statistics

$Q_n(P)$ is the unknown joint distribution of T_n.

$\mathcal{S}_0 \equiv \{j : H_{0j} \text{ is true}\}$ is the set of true null hypotheses.

$\mathcal{S}_n \subset \{1, \ldots, m\}$ indicates the set of hypotheses for which one rejects the null hypotheses based on some MTP.

V_n is the number of falsely rejected null hypotheses in the multiple testing procedure $\mathcal{S}_n$.

$\theta(F_{V_n})$ is the type I error rate, which is a function of the distribution of the number of false positives (F_{V_n}), one wishes to control with the MTP.

$S(T_n, Q, \alpha) = \{j : T_n(j) > c_j(Q, \alpha)\}$ is a MTP based on chosen cut-off rule $c(Q, \alpha)$, where α is the desired error rate, and Q is the proposed null distribution.

Q_{0j} is a marginal null distribution for the jth test statistic that dominates the true marginal distribution, Q_{nj}, of the test-statistic for $j \in \mathcal{S}_0$ (true nulls).

$\tilde{T}_n(j) = Q_{0j}^{-1} Q_{nj}(T_n(j))$ is the proposed finite sample null distribution of $T_n(j)$.

$Q_{0n}(P)$ is the finite sample joint null distribution dominating the true distribution $Q_n(P)$ of T_n, and $Q_0(P)$ denotes the asymptotic limit distribution of $Q_{0n}(P)$.

$Q_{nj}^{\#}$ is the marginal bootstrap distribution of $T_n(j)$, and $T_n^{\#}(j)$ is the test-statistic calculated on a bootstrap sample.

$\tilde{T}_n^{\#}(j) = Q_{0j}^{-1} Q_{nj}^{\#}(T_n^{\#}(j))$ is the bootstrap analogue of $\tilde{T}_n(j)$.

DOI: 10.2202/1544-6115.1199

References

Sandrine Dudoit, Mark J. van der Laan, and Katherine S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art13`. Article 13.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96(456):1151:1160, 2001.

C.R. Genovese and L. Wasserman. Exceedance Control of the False Discovery Proportion. Technical Report 762, Department of Statistics, Carnegie Mellon University, July 2003. URL `http://www.stat.cmu.edu/cmu-stats`.

R. D. Gill. Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.*, 16(2):97–128, 1989. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author.

E.L. Lehmann and J.P Romano. Generalizations of the Family-wise Error Rate. *Annals of Statistics*, 33:1138:1154, 2005.

Katherine S. Pollard and Mark J. van der Laan. Resampling-based Multiple Testing: Asymptotic Control of Type I error and Applications to Gene Expression Data. Technical Report 121, Division of Biostatistics, University of California, Berkeley, June 2003. URL `http://www.bepress.com/ucbbiostat/paper121`.

D.R. Skibola, M.T. Smith, P.M. Bracci, A.E. Hubbard, L. Agana, S. Chi, and E.A. Holly. Polymorphisms in ghrelin and neuropeptide y genes are associated with non-hodgkin lymphoma. *Scand. J. Statist.*, 14(5):1251–6, 2005.

Mark J. van der Laan, Merrill D. Birkner, and Alan E. Hubbard. Empirical bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2004a. URL `http://www.bepress.com/sagmb/vol4/iss1/art29`. Article 29.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. Technical Report 1, 2004b. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Multiple Testing. Part II. Step-Down Procedures for Control of the Family-Wise Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004c. URL `http://www.bepress.com/sagmb/vol3/iss1/art14`. Article 14.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Multiple Testing. Part III. Procedures for Control of the Generalized Family-Wise Error Rate and Proportion of False Positives. Technical Report 141, Division of Biostatistics, University of California, Berkeley, January 2004d. URL `http://www.bepress.com/ucbbiostat/paper141`.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

P. H. Westfall and S. S. Young. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley and Sons, 1993.

Z. Yu and M.J. van der Laan. Construction of counterfactuals and the g-computation formula. Technical report, Division of Biostatistics, UC Berkeley, 2002.

DOI: 10.2202/1544-6115.1199