Statistical Methods in Medical Research 2007; 1–42

A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion

Alessio Farcomeni Università di Roma 'La Sapienza', Roma, Italy

In the last decade a growing amount of statistical research has been devoted to multiple testing, motivated by a variety of applications in medicine, bioinformatics, genomics, brain imaging, and so on. Research in this area is focused on developing powerful procedures even when the number of tests is very large. This paper attempts to review research in modern multiple hypothesis testing with particular attention to the false discovery proportion, loosely defined as the number of false rejections divided by the number of rejections. We review the main ideas, stepwise and augmentation procedures; and resampling based testing. We also discuss the problem of dependence among the test statistics. Simulations make a comparison between the procedures and with Bayesian methods. We illustrate the procedures in applications in DNA microarray data analysis. Finally, few possibilities for further research are highlighted.

1 Motivation

In many areas of application of statistics, in particular in bioinformatics, conclusions are drawn by simultaneous testing of a large number of hypotheses. In these high-dimensional situations common single inference approaches are well known to fail, leaving open the problem of making a small number of false discoveries by controlling a suitable error rate, and maximizing the power of each test at the same time. Such problem of simultaneous inference is usually referred to as *multiple testing*. Applications in multiple testing include identifying neuronal activity in the living brain or the identification of differentially expressed genes in DNA microarray experiments.[1–10] For a review of multiple testing methods in the context of microarray data analysis, see Ref. [11] and Ref. [12] for an excellent review of genomics and statistical challenges in genomics. Among the other possible applications, there are general medicine,[13] pharmacology,[14] epidemiology,[15] psychometrics[16] and even marketing.[17]

Moreover, multiple tests can be used as a key part of statistical procedures, like variable selection,[18,19] item-response modeling,[20] structural equation modeling,[21] decision trees,[22] wavelet thresholding,[23,24] and so on.

Many of these applications have arisen recently, posing new kinds of multiplicity problems and stimulating a tremendous interest and fast developments in multiple hypothesis

Address for correspondence: Alessio Farcomeni, Università di Roma 'La Sapienza', Piazzale Aldo Moro, 5, 00185, Roma, Italy. Email: alessio.farcomeni@uniroma1.it

10.1177/0962280206079046

testing. This paper attempts to review research in modern multiple hypothesis testing, with particular attention to the false discovery proportion (FDP), the basis for some of the newly introduced error rates. The FDP can be defined to be the number of false positives divided by the number of rejections.

In this paper we will mainly consider distribution-free methods for testing on a simple hypotheses with the use of significance levels (p-values).

The paper is organized as follows: Section 2 will describe the multiple testing framework, and introduce the most popular Type I error rates. Section 3 will introduce the main concepts about multiple testing procedures (MTPs). Section 4 will review some MTPs controlling classical and modern error rates. Section 5 will give a general comparison of the procedures through simulations, and show the performance of (not corrected) Bayesian procedures in terms of classical error measures. Section 6 will review extensions under dependence among the test statistics. In Section 7, we show some real life applications and finally, Section 8 will conclude with a brief discussion, and point out some possibilities for further research in this area.

2 The multiple hypothesis framework

Consider a multiple testing situation in which m tests are being performed. Suppose M_0 of the m hypotheses are true, and M_1 are false. Table 1 shows the possible outcomes in testing m hypotheses: we denote with R the number of rejections, with $N_{0|1}$ and $N_{1|0}$ the exact (unknown) number of errors made after testing; and with $N_{1|1}$ and $N_{0|0}$ the number of correctly rejected and correctly retained null hypotheses. The number of rejected hypotheses R is random, and consequently all $N_{i|j}$; while M_0 and M_1 can either be considered as random or just not observable, depending on the specific application. For the time being we assume all the test statistics are independent, and will discuss below generalizations to dependent test statistics.

In the usual (single) test setting, one controls the probability of false rejection (Type I error) while looking for a procedure that possibly minimizes the probability of observing a false negative (Type II error).

In the multiple case, despite each uncorrected level α test falsely rejects the null hypothesis with small probability (namely, α), as m increases the number of false positives can explode. For instance, if $m = 30\,000$ true null hypotheses are simultaneously tested at level $\alpha = 0.05$, around $R = N_{1|0} = 1500$ false discoveries are expected. The consequences of so high a number of false discoveries in real applications would usually be extremely deleterious.

Table 1 Outcomes in testing m hypotheses

	H_0 not rejected	H_0 rejected	Total
H_0 True	$N_{0\|0}$	$N_{1\|0}$	M_0
H_0 False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$m - R$	R	m

From a different point of view it can be said that a p-value around, for instance, 0.05 is unlikely to be correspondent to a true discovery, since it is very likely under the null hypothesis that such a small p-value will occur when many are computed at once.

Corrections arise from the control of specific Type I error measures, and there are a variety of functions of the counts of false positives $N_{1|0}$ that can serve as possible generalizations of the probability of Type I error. Control of the chosen Type I error rate can be loosely defined to be achieved when the error rate is bounded above by a pre-specified $\alpha \in (0, 1)$. A more detailed discussion is given below.

The most classical multiple Type I error rate is based only on the distribution of $N_{1|0}$, that is, on what happens for the tests corresponding to the true null hypotheses:

- *Family-wise error rate* (FWER), the probability of a least one Type I error:

$$\text{FWER} = \Pr(N_{1|0} \geq 1) \tag{1}$$

Here and in what follows, unless stated otherwise probability and expectations are computed conditionally on the true parameter configuration, that is, on which and how many hypotheses are true. In common approaches equal importance is given to each hypothesis. Holm[25] discusses adjustments of certain FWER controlling procedures when different importance is explicitly given to the hypotheses, by weighting; and[26] generalize it to enhance power in high-dimensional situations. See also Section 3.4.

Many modern Type I error rates are based on the FDP; defined to be the proportion of erroneously rejected hypotheses, if any:

$$\text{FDP} = \begin{cases} \dfrac{N_{1|0}}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} \tag{2}$$

The FDP is then based also on the distribution of R, that is, on what happens for the hypotheses for which H_0 is false. Benjamini and Hochberg[27] propose to control the expectation of the FDP, commonly refereed to as False Discovery Rate (FDR). The first to consider this error measure was probably Seeger[28] who advocated control of FWE but additional checking of the proportion of false nulls. Dudoit *et al.*[29] and independently Genovese and Wasserman[30] along similar lines propose to control the tail probability of the FDP (tFDP(c)). This error measure is sometimes referred to as false discovery exceedance (FDX):

- *FDR*, expected proportion of Type I errors:

$$\text{FDR} = E[\text{FDP}] \tag{3}$$

- *FDX*, tail probability of the FDP:

$$\text{tFDP}(c) = \Pr(\text{FDP} > c) \tag{4}$$

Control (in expectation or in the tail) of the FDP is justified by the idea that any researcher is prepared to bear a higher number of Type I errors when more rejections are made. We note that the only assumptions needed to control the FDR or FDX are usually related to the dependence among the test statistics.

Further generalizations of the FWER and FDR are proposed in Refs. [31–34]. Storey[32] for instance introduced the positive FDR, defined as $\text{pFDR} = E[\text{FDP}|R > 0]$. Control of this error measure is more appropriate when the probability of making no rejections is high, so that FDR control may be misleading; and can moreover lead to more powerful multiple testing procedures in certain situations. Note that for any number of rejected hypotheses $\text{FDR} \leq \text{pFDR}$. Storey[32] suggested how to estimate and thus control pFDR fixed rejection region, and introduce the q-value, a pFDR analogue of the p-value. An interpretation of the pFDR and q-value as Bayesian posterior probabilities is in Ref. [35], that also shows connections to classification theory. A discussion of weighted FDR controlling procedures, instead in Refs. [36] and [37], also show how to give difference importance to each hypothesis, and also how to enhance power by weighting.

It is straightforward to see that FDR and FDX control is also a weak control on the FWER, in the sense that FWER is controlled if all the null hypotheses are true, and that $\text{tFDP}(0) = \text{FWER}$.

The introduction of FDP-based error measures was motivated by some modern applications, in which the number tests can be very large. In these settings, FWER controlling procedures tend to become conservative and finally lead to rejection of a very limited number of hypotheses, if any. Conversely, FWER control is more desirable when the number of tests is small, so that a good number of rejections can be made, and all can be trusted to be true findings.

A general comparison of the error rates may be summarized by the inequalities

$$\begin{aligned}\frac{E[N_{1|0}]}{m} &\leq \min(\text{FDR}, \text{FDX}) \leq \max(\text{FDR}, \text{FDX}) \\ &\leq \text{FWE} \leq E[N_{1|0}]\end{aligned}$$

that are straightforward to prove. The last term $E[N_{1|0}]$ is sometimes referred to as *per family* Type I error rate, and the first $E[N_{1|0}]/m$ as *per comparison* error rate; as in fact $E[N_{1|0}]$ is the expected number of type I errors and $E[N_{1|0}]/m$ is the expected marginal probability of erroneously rejecting a given hypothesis.

Genovese and Wasserman[38] and independently Sarkar[39] generalize the concept of Type II error to the multiple case by introducing the False Negatives Rate (FNR), the dual of the FDR, defined as:

$$E\left[\frac{N_{0|1}}{m - R + 1_{(m-R)=0}}\right] \tag{5}$$

Sarkar[39] also introduces the concept of unbiasedness of an FDR-controlling procedure, as a procedure satisfying the inequality $\text{FDR} + \text{FNR} \leq 1$.

2.1 Relationship between FDR and FDX control

FDX and FDR are closely related, being functionals of the same random variable, namely, the FDP.

It is straightforward to see[30] that in general if tFDP(c) is controlled at level α, then FDR is controlled at level $c + (1 - c)\alpha$. In Section 5.2 we will give some ideas on how to choose c in order to use a tFDP(c) controlling procedure to suitably control the FDR.

A partial converse is given by an application of Markov inequality, which yields: if FDR $< \alpha$ then tFDP(c) $< \alpha/c$.

Moreover, note that FDR $= E[FDP] = \int_0^1 \text{tFDP}(c)\, dc$, that is, FDR control is a control on the average tFDP(c) (with respect to Lebesgue measure). Following this statement, we can apply the mean value theorem and prove that at least asymptotically there exist $\xi \in [0, 1]$ such that tFDP(ξ) $= FDR$. That is, if FDR $\leq \alpha$, there exist $\xi \in [0, 1]$ for which tFDP(c) $\leq \alpha$ for any $c > \xi$. Simulations (Section 5.1) suggest that this 'mean value' ξ is often not small, as FDR control rarely implies FDX control with the typical choice $c = 0.1$; while the converse is often true.

2.2 The statistical model

In this section we will formalize the framework for multiple testing. Let $X_1, \ldots, X_n$ be n random vectors in R^m: $X_i = (X_{ij}: j = 1, \ldots, m)$. To fix the ideas suppose we observe n replicates of m variables, arising from a certain multivariate distribution $\mathcal{P}$, depending on unknown parameters. The n replicates are combined to compute test statistics, finally arising a vector of m p-values.

The sample size n is typically much smaller than m: for instance, one can measure the expression level of thousands of genes in less than one hundred patients, together with biological covariates and risk factors. This obviously complicates the shape and definition of $\mathcal{P}$.

2.2.1 Parameters

We define a parameter to be a function of the unknown distribution $\mathcal{P}$. Parameters of interest typically include means, differences in means, variances, ratios of variances, regression coefficients, and so on.

2.2.2 Null hypotheses

We claim that a null hypothesis is true if a specified model holds, while the alternative is true if the data is not distributed according to the specified model. In this paper, anyway, we focus on testing simple one-parameter hypotheses, that is, the null hypothesis specifies a value for a parameter of interest. For instance, testing to see whether a difference in means is zero, like in paired T-testing.

Let S_0 be the set of all true null hypotheses. The goal of a multiple testing procedure is the accurate estimation of S_0, while probabilistically controlling a pre-specified function of the number of false rejections $N_{1|0} = \{S_0 \cap \widehat{S}_0^c\}$, that is, a Type I error rate.

When $m = 1$, we have a classical statistical test, which can be seen as a procedure partitioning the sample space in two: a subset $\mathcal{X}_1 \subseteq \mathcal{X}^n$ such that observing a realization of $X_1, \ldots, X_n$ in $\mathcal{X}_1$ leads to rejection of the single null hypothesis; and the complementary subset, leading to failure of rejection.

This idea can be generalized to the $m > 1$ case: a multiple test is a procedure partitioning the sample space in 2^m subsets, each corresponding to one of the 2^m possible estimates of S_0, that is, to a set of rejected hypotheses.

2.2.3 *Test statistics and p-values*

Partitioning of the sample space is made through a vector of *m test statistics*, $T_n = (T_n(j): j = 1, \ldots, m)$, that are functions of the data $X_1, \ldots, X_n$. We define the j-th p-value to be

$$p_j = \Pr(|T_n(j)| > |t_n(j)| \,|\text{the } j\text{-th null is true}) \tag{6}$$

where $t_n(j)$ is the observed value of the test statistic $T_n(j)$. Throughout we adopt the notation $p_{(j)}$ to denote the j-th ordered statistic of the vector of p-values, with $p_{(0)} = 0$ and $p_{(m+1)} = 1$. The p-value p_j is again a random variable, when the null hypothesis is simple, and the distribution of the test statistics is continuous and known, it is uniformly distributed on [0, 1] under the null; while it is stochastically dominated by the uniform if the distribution is discrete. In certain practical situations, the probability in Equation (6) cannot be directly computed, and an estimate of the p-value must be computed, often by a *resampling* method. The estimates are not in general uniformly distributed, and these are what many methods use. For further discussion on estimation of p-values see Section 3.2.

It is equivalent to work with test statistics or p-values. Providing corrections for p-values is usually more convenient since they are bounded and their distribution under the null is invariant with respect to the data generating distribution.

2.2.4 *Families*

An important issue in multiple testing is what set of hypotheses to consider in the computation of the error rate. In many situations the hypotheses can be divided in *families*, groups of different types and usually of limited size. In large surveys and multifactorial experiments families usually arise naturally by groups of variables, time periods, phases of the study, category of the endpoints, etc. That is to say, questions form natural units indexed by a time point and kind of issue. For real life examples see Ref. [40] (Chapter 7), and Section 7.1.

Westfall and Young[40] note also that it may be sensible to consider as a whole family the tests computed for a publication.

Issues in deciding a family are considered moreover in Refs. [41,42], while Ahmed [43] suggests strategies that involve the consideration of several different families in large surveys.

The choice of a family is in part subjective, and different choices may lead to sensibly different results. The effect is particularly strong when the majority of the tested nulls are in fact true. To avoid the possibility of data snooping, families should be defined before seeing the data whenever possible, and a clear description and justification of the grouping considered should be reported together with the results.

In other cases there is no basis for a natural division into groups, and all the hypotheses involved must be considered as a single family. In bioinformatics applications for instance the definition of a family is not an issue, as genes, neurons, and so on will always be

considered together. In these applications the number of hypotheses tested together can be very large, magnifying the problems related with the multiplicity of the comparisons.

3 General properties of multiple testing procedures

A MTP produces a set $S_{\text{MTP}} = \{j\colon p_j \leq T\}$ for the (random) cut-off T of rejected hypotheses, which is an estimation of the set S_0^c of false null hypotheses. Since p-values (and test-statistics) live in $\mathcal{R}$, the complex procedure of partitioning an high-dimensional space in 2^m slices reduces to fixing a cut-off T such that all the hypotheses corresponding to p-values below T are rejected and the others retained. Control of Type I error rates substantially reduces to just showing how to choose such T, that in general (but not necessarily) will be below the single inference cut-off α. Note that even if in *stepwise* methods there is comparison of each p-value p_j with a different constant, α_j, we still can set T as the biggest rejected p-value. In that case T will be random by definition.

For a fixed MTP, the set S_{MTP} depends on:

- The data, often only through the vector of p-values $p_1, \ldots, p_m$
- A level α, that is, an upper bound for an appropriate Type I error rate.

Some authors, for example,[11,40,44,45] prefer to leave the cut-off T fixed to α and introduce the concept of *adjusted p-value*, a level of significance related to the entire MTP. In practice, this involves defining a new p-value $\tilde{p}_j$, which will be a function (often, a scale transformation) of the old p_j. More formally, define the adjusted p-value

$$\tilde{p}_j = \inf\{\alpha\colon \text{The } j\text{-th hypothesis is rejected by the procedure with nominal error rate } \alpha\}.$$

For instance, if each of the hypotheses is tested with the Bonferroni correction at level α/m, the adjusted p-value for each hypothesis is just $\tilde{p}_j = mp_j$. Then, it is possible to define $S'_{MTP} = \{j\colon \tilde{p}_j < \alpha\}$. Strategies to compute adjusted p-values when complex multiple testing procedures are used are given in Ref. [44]. It can be shown that, for each MTP procedure, it is perfectly equivalent to consider adjustment of threshold or of p-values. The additional computation of adjusted p-values is often very useful in practice, since adjusted p-values are easily interpretable, directly comparable with other multiple testing experiments, and a more intuitive choice of the level of the test is possible.

3.1 Types of multiple testing procedures

MTPs are usually categorized as:

- **One-step**: In one-step procedures, all p-values are compared to a predetermined cut-off, usually only a function of α and m, with no dependence on the data.
- **Step-down**: In step-down procedures, each p-value is compared with a cut-off dependent on its rank. The p-values are examined in order, from smallest to largest. At each step, the p-value is rejected if smaller than its cut-off. Once a p-value is greater than its cut-off, it is not rejected together with all the higher ones.
- **Step-up**: Step-up procedures are similar to step-down procedures. p-values are examined from the largest to the smallest. At each step, the p-value is not rejected if

larger than its cut-off. Once a p-value is found to be significant, it is rejected together with all the smaller ones.

In step-up and step-down methods it is not uncommon to talk about 'step-up/step-down' constants α_j. This is referring to the threshold with which the j-th ordered p-value is compared. Hochberg and Tamhane[46] note that if a step-up and a step-down procedure are based on the same constants, the step-up version will reject at least the same number of hypotheses, thus being at least as powerful as the step-down version. Finner and Roters[47] provide the distribution of the number of false discoveries for stepwise methods. Finally, note that setting $J = \max\{j\colon p_j \leq \alpha_j\}$, the index of the biggest p-value below the step-up constant α_j in step-up methods, and $J = \min\{j\colon p_j > \alpha_j\} - 1$, the index of the largest p-value smaller than the smallest p-value above the step-down constant α_j; one can refer to a single random cut-off $T = p_{(J)}$.

Multiple testing procedures may also be *augmentation* based. Augmentation procedures proceed iteratively, first rejecting a certain number of hypotheses and then rejecting an additional number chosen as function of the number rejected at the first step.

3.2 Resampling based testing

Resampling based testing relies mainly on the idea that the data can often be resampled in a way that reflects the null hypothesis. As an example, consider the case of comparing a quantitative response between two groups. Under the hypothesis of no difference between groups, observations arise just from a random assignment to the groups. When data are resampled and randomly assigned to the groups, one can expect to see no systematic differences. If the original test statistic is unusual with respect to the resampling distribution, then null hypothesis and data are in conflict. This conflict can be measured, as usual, with a (resampling-based) p-value: the proportion of times the resampled statistic is at least as extreme as the original observed test statistic.

On can resample each variable and determine individual p-values independently. In our context it is anyway more frequent that the whole multivariate distribution is used, for instance by resampling vectors. This is particularly useful, since the joint distribution can be explored without explicitely modelling dependence. Resampling procedures achieve (exact or asymptotic) error control under general dependence, and further they will be less conservative because the information provided by actual dependence in the data will be exploited.

3.2.1 Permutation methods

In permutation testing[48] the data is sampled without replacement: at each iteration of the algorithm the observations are simply randomly shuffled. Note that this implies that certain statistics (for instance, the overall mean in the previous single test example) are constant with respect to resampling; and this can be sometimes used to speed up the algorithms. For an example, see for instance[40] (Chapter 5.3).

Permutation methods can be seen in many settings as a *nonparametric* and *conditionally exact* approach to testing; in the sense that the permutation p-values will not depend on the underlying population distribution and that, under the permutation null hypothesis that all permutations are equally likely, they are exact.

It is intuitive that the *p*-values, unconditionally, need not be uniformly distributed. Nevertheless, they are conservative: their distribution is stochastically larger than the uniform for any finite *n*. For this reason, the *p*-values obtained are *valid*, in the sense that they lead to error rate control in finite samples.

Permutation methods anyway require certain assumptions, in particular exchangeability under the null hypothesis, which is not guaranteed to hold in a few relevant cases; and furthermore exhaustive permutation is sometimes not feasible, so that only a random sample of permutations are considered and only approximate Type I error control is guaranteed.

3.2.2 Bootstrap methods

In bootstrap testing[40,49] the data is sampled with replacement.

The main idea is that while the dependency structure can be preserved (by resampling the multivariate distribution), it is possible to resample with strategies that essentially remove the systematic effects from the data, obtaining an estimate of the null distribution: it often happens that bootstrap samples are re-centered (while permutation samples are not).

Bootstrap *p*-values can either be conservative or anti-conservative. In the second case, the Type I error rate is inflated; and only *asymptotic* control is guaranteed. The finite sample properties of the bootstrap still need much investigation. On the other hand bootstrap is more general than permutation since exchangeability is not required, and furthermore more complex models can be accommodated with bootstrap strategies.

3.2.3 Discussion about resampling methods

In general resampling based testing provides powerful procedures that are also often valid under general dependence. Furthermore, complex situations in which other solutions are not available can be often tackled with resampling methods.

Main drawbacks are that 1) methods may have poor properties in small sample situations, in our context expecially whenever $n \ll m$; and 2) they are often time consuming and computationally intensive. The latter problem, even in the presence of powerful computing facilities, may be considerable in large *m* situations.

In cases in which the number of samples is small compared to the number of tests it may even be inappropriate or not useful to apply resampling methods. As a referee pointed out, it is known for instance that if *n* is sufficiently small relative to *m* the permutation methods will never reject, thus having zero power. A similar low power problem is likely to happen for bootstrap methods:[50] argue in a simulation study that when the sample sizes are small, bootstrap can lead to testing procedures that have much lower power than permutation tests.

A further problem is given by a lack of generality: usually ad-hoc strategies need to be derived in order to implement a bootstrap or permutation method. Additional ad-hoc considerations are often useful in order to speed up the methods and increase precision, expecially when performing many tests at once.

As pointed out by Pollard, van der Laan,[51] there are many possibilities for estimating the null distribution through a given resampling strategy. They suggest projecting the

true test statistic distribution onto the space of null (mean zero) distributions by bootstrapping centered test statistics. They prove that bootstrapping centered test statistics provides asymptotic strong control of the desired error measure. An illustration of this methodology in a resampling-based method for controlling the FDX will be given below. Results of Ref. [51] are further discussed in Ref. [45], who show a general characterization of null distributions leading to asymptotic control of the error measures.

Ge *et al*.[52] review resampling based multiple testing in the setting of Microarray data analysis. Some examples of resampling based multiple testing procedures are given in Section 4.

3.3 Types of Error control

Let now $\mathrm{ER}_m(P)$ be any error rate for a fixed m and conditionally on distribution P for the data. The control of the error rate can be categorized as:

- **Weak control**: there is weak control of the Type I error rate $\mathrm{ER}_m(P)$ if there is finite sample control of $\mathrm{ER}_m(P_0)$, where P_0 is the distribution that would have generated the data if all the null hypotheses were true (the so called 'complete null').
- **Strong control**: there is strong control of the Type I error rate $\mathrm{ER}_m(P)$ if there is finite sample control of $\max_{I\subseteq\{1,\ldots,m\}} \mathrm{ER}_m(P_I)$, where P_I is the distribution that would have generated the data if the null hypotheses indexed in I were true and the other false. That is, strong control implies that the Type I error measure is bounded above no matter the configuration of true and false hypotheses.
- **Exact control**: The control of $\mathrm{ER}_m(P)$, that is, the achievement of $\mathrm{ER}_m(P) \leq \alpha$, where P is the true parameter configuration.

We do not want to control errors under configurations that are not the true ones, possibly losing power. Weak and strong control are nevertheless introduced since it may not be doable to work with $\mathrm{ER}_m(P)$, with unknown P. Obviously, strong control implies weak control; and if attaining weak control, one hopes that the error rate under the true distribution, as often happens, is bounded by the error rate under the complete null. The same idea applies to strong control, since it is intuitive that the error rate under the true distribution is bounded by the maximal error rate under all possible configurations, $\max_{I\subseteq\{1,\ldots,m\}} \mathrm{ER}_m(P_I)$. Moreover, the results of a weak controlling procedure can be used only to imply that there is some false hypothesis among the m, but do not allow the researcher say *which* hypotheses are false. For this reason, weak FWE controlling procedures are sometimes referred to as 'omnibus' tests.

A multiple testing procedure provides *asymptotic exact control* of a pre-specified Type I error rate at level α if $\limsup_n \mathrm{ER}_m(P) \leq \alpha$, where P is the true distribution. Asymptotic weak and strong control can be defined similarly.

While asymptotic is usually meant in n, there also are some works which consider the possibility of a growing number of tests, like,[47,53,54] and some references therein. These results are useful for applications in which $m \gg n$.

Note also that weak and strong control of pFDR is not possible, since by definition if $M_0 = m$ then $\mathrm{pFDR} = 1$.

We also introduce the idea of *subset pivotality*, as defined in Condition 2.1 of Ref. [40]. Subset pivotality holds for the vector of p-values if, for any $k \in 1, \ldots, m$, it happens that $\{p_{j_1}, \ldots, p_{j_k}\} \stackrel{d}{=} \{p_{j'_1}, \ldots, p_{j'_k}\}$. In words, the multivariate distribution of any subset of p-values is unaffected by hypotheses not considered in the subset, and in particular by the truth or falsehood. If subset pivotality holds, strong control is implied by weak control of the Type I error rate. This is a technical condition which is often trivially satisfied, for instance by the existence of no logical constraint making impossible a particular combination of null and false hypotheses. A situation in which subset pivotality fails is when testing on the elements of a correlation matrix.[51,55]

In this paper we focus on distribution free methods, that is, procedures that control the chosen Type I error rate under any distribution for the p-values under the alternative $p_i|H_i = 1$, which we denote by $F(\cdot)$; and any configuration of the true and false nulls.

There are also many methods that are based on parametric assumptions, which dominated the early literature. Some of them can, for instance, be found in Ref. [46].

3.4 A special case: multiple endpoints

While the aim of the paper is to review general procedures for multiple testing correction, it is interesting to mention the multiplicity problems arised by the evaluation of multiple endpoints in clinical studies.

In clinical trials the evaluation of multiple endpoints is one of the most common problems arising multiplicity issues. While the multiple testing procedures we review are directly applicable in this setting, there often is a hierarchy among the tests that allows for further considerations. In many cases in fact one is able to classify the tests in groups of primary and secondary (and even thertiary) endpoints. The number of primary endpoints is usually limited, and often a single primary endpoint is encountered (for instance, reduced mortality); while the number of secondary endpoints (for instance, side effects) may be very large. Furthermore, even if there still is the necessity to be very careful about conclusions, the researcher usually expects that a large proportion of the primary endpoints is in fact significant, while a large proportion of the secondary endpoints may in fact be non-significant. For a general discussion about clinical trials with multiple outcomes, see Refs [56] and [57].

There are many possible approaches to tackle this case.

The easiest approach would be to ignore the hierarchy in the definition of the endpoints, and control a suitably chosen Type I error rate on the whole set of tests.

In the case of a single primary endpoint, furthermore, it is a common approach not to apply any correction for multiplicity. However, while the results for the primary endpoint are reliable, for what concerns the secondary endpoints the tests would only have an exploratory value.[57,58] In particular, Chi[58] suggests that, without significance of the primary endpoint, no significance can be claimed for any secondary endpoint independently on the correponding p-values magnitude.

A further more elegant possibility is to do error allocation[59,60]. See also Ref. [36]. Suppose that the nominal FWER level for the entire experiment is set to α_E, say $\alpha_E = 0.05$. Under independence (or positive orthant dependence, see Section 6), this error rate

can be differentially allocated between primary and secondary endpoints according to the formula:

$$\alpha_E = 1 - (1 - \alpha_P)(1 - \alpha_S) \tag{7}$$

where α_P is the error rate for primary endpoints and α_S is the error rate for secondary endpoints. The formula comes from simple probability considerations. Suppose there are n_S secondary endpoints. The overall error rate for secondary endpoints can then be allocated to each endpoint according to the formula:

$$\alpha_S = 1 - \prod_{i=1}^{n_S}(1 - \alpha_{S_i})$$

and similarly if there is more than one primary endpoint. Of course the researcher can not choose any value for α_E, α_P and α_S, since any two will force the other, and similarly for the individual endpoint error rates α_{S_i}. For instance, choosing $\alpha_E = 0.05$ and $\alpha_P = 0.03$ will force $\alpha_S = 0.021$.

Under general dependence the previous formulas do not hold, and should be replaced with their Bonferroni equivalents:

$$\alpha_E = \alpha_P + \alpha_S \tag{8}$$

and for instance

$$\alpha_S = \sum_{i=1}^{n_S}\alpha_{S_i}$$

Such Bonferroni values are in many situations only slightly more conservative than the ones relying on assumptions on the dependence.

Note that of course α_P or α_S must be in all cases set to a smaller value than the experimental rate α_E.

There are many remarks related to error allocation methods. First, the entire procedure must be performed before the experiment. In fact, there are several different ways to allocate an overall error α_E, which can potentially lead to very different conclusions. Allocating the error rate after seeing the data is data snooping and leads to inflated error rates, possibly higher than the nominal levels. This also implies that the strategy can be applied only if the endpoints are prospectively determined. Moreover, while there is an historical justification for the use of overall errors of 5%, there is no agreed convenction for the allocation of such errors, so that when doing error allocation the choices should be clearly indicated, justified, and also the non-significant p-values should be reported. A choice that is usually easily justifiable is to set constant the error rate within each level, that is, $\alpha_{S_1} = \alpha_{S_2} = \cdots = \alpha_{S_{ns}}$ and similarly for the primary endpoints; unless there is a clear reason. We also note that the allocation procedure is conservative under dependence of the test statistics.

In certain cases it may be justifiable the division of the hierarchy in separate families. In that case a different error rate may be controlled on each. FWER may be more

appropriate for the primary endpoints and FDR/FDX for the secondary endpoints, due to the previous considerations. Note however that endpoints are naturally depedent, so that one should use procedures valid under general dependence.

There are also approaches that use omnibus multivariate tests based on parametric assumptions (like multivariate analysis of variance, Hotelling's T-test, and similar). See for instance.[61–64] In particular, O'Brien [61] develop a least-squares test statistic that is more powerful than Hotelling's test statistic when the endpoints are positively correlated, under the assumption of normality; and Pocock *et al*.[62] generalize the idea to test statistics that are asymptotically normal. These ideas are then applied to survival analysis by Wei and Lachin[65] and Wei *et al*. [66] Recall anyway that omnibus tests provide only weak control of the FWE and then when using this approach it is not possible to drive conclusions about the individual endpoints. Nevertheless, in certain cases omnibus testing may be more useful for other purposes:[67] note that Bonferroni and stepdown tests strongly controlling the FWE are more suitable for detecting one highly significant difference, while omnibus tests can be more powerful in rejecting the global null when each individual test statistic is barely significant.

A further common approach is to compute a single aggregate test statistic (for instance, time to first event or a summary score) for each level of endpoints. This is useful if one is not interested in the results of individual endpoints.

Apart from primary and secondary endpoints in clinical studies, there are other relevant settings for which special adjustment procedures and specific considerations can be made. We mention for instance simultaneous comparison of more than two groups.[46]

4 Multiple testing procedures

4.1 General ideas behind a multiple testing procedure

Before reviewing the procedures that control each Type I error rate, we summarize the general ideas:

- We want to fix a cut-off T such that the error rate is at most equal to a prespecified $\alpha \in [0, 1]$.
- This cut-off should be as high as possible, provided the specified error rate is controlled. The higher T, the more tests are rejected and the more powerful the procedure.
- Consequently, if two procedures control the same error rate, we prefer the one achieving a smaller Type II error rate.

Multiple testing procedures aim at a balance between false positives (given by larger Ts) and false negatives (given by smaller Ts). We follow here the principle that, as long as the Type I error rate is controlled at the desired level, we prefer to make more false positives in order to have less false negatives. To simplify the exposure, we will say 'reject $p_{(j)}$ such that...' to indicate 'reject the hypotheses corresponding to $p_{(j)}$ such that'.

All the procedures we will review provide (finite sample or asymptotic) strong control of the considered error rate, unless stated otherwise.

4.2 Procedures controlling the FWER

In this subsection, we briefly review procedures to control the FWER, as defined in Equation (1). More details can be found in the books by Westfall, Young[40] or Hochberg and Tamhane.[46]

Bonferroni Bonferroni correction is a one-step method at level $T = \alpha/m$.

Step-down Holm[25] propose to improve on Bonferroni by using the step-down constant $\alpha_j = \alpha/(m - j + 1)$.

Step-up Hochberg[68] proves the same constant of Holm can be used in a (more powerful) step-up method.

Step-down minP Let $F_{r,\alpha}(\cdot)$ indicate the α percentile of the distribution of the minimum of the last r p-values. The 'Step-down minP' procedure fixes a step-down constant $\alpha_j = F_{j,\alpha}(p_{(m-j+1)}, \ldots, p_{(m)})$.

One-step Sidak The one-step Sidak procedure consists in controlling each test at a level $1 - \sqrt[m]{1-\alpha}$.

Step-down Sidak A step-down version of Sidak correction consists in using the step-down constant $\alpha_j = 1 - \sqrt[m-j+1]{1-\alpha}$.

A classical procedure to control the FWER is the Bonferroni correction, which is a one-step method fixing $T = \alpha/m$. Hence, one would reject only the hypotheses for which $p_j \leq \alpha/m$. It is easily seen that this controls the FWER under arbitrary dependence.

Step-up Hochberg can be more powerful than step-down Holm, but as we will point out later it is less robust with respect to dependence.

For the step-down minP procedure, quantiles arise from the distribution of the minima of the last k p-values. The minP procedure was first proposed in Ref. [40], who propose to estimate $F_{j,\alpha}(\cdot)$ through resampling, and give a double-permutation algorithm. This algorithm is slow since it involves resampling the null distribution within a resampling scheme. Pesarin[48] and independently Ge *et al.*[52] suggest a much quicker permutation algorithm which we briefly describe:

1. Set $i = m$
2. For the hypothesis corresponding to the i-th ordered p-value, compute B permutation p-values $p_{i,1}, \ldots, p_{i,B}$.
3. Compute the successive minima $q_{i,b} = \min(q_{i+1,b}, p_{i,b})$, with $q_{m+1,b} = 1$, and estimate the i-th adjusted p-value as $\tilde{p}_i = \sum_b 1_{q_{i,b} \leq p_{(i)}}/B$. Set $i := i - 1$.
4. If $i > 0$, go to step 2. If $i = 0$, enforce monotonicity of the estimated p-values: $\tilde{p}_i = \max(\tilde{p}_{i-1}, \tilde{p}_i)$, $i = 2, \ldots, m$. Reject the hypotheses for which the estimated adjusted p-values are below α.

This algorithm is particularly useful when m is large, as resampling is done just once and iterations are particularly quick. The main idea is that it is possible to proceed one hypothesis at a time, unlike the classical algorithm. The procedures starts with an estimate of the least significant adjusted p-value. It computes B permutation testing $p_{(m),1}, \ldots, p_{(m),B}$, and estimates the least significant p-value as the proportion of resampled maxima above the observed $p_{(m)}$. Each $p_{(m),1}$ comes from a permutation of the test statistics for the (m)-th hypothesis. The procedure is then repeated for the second least significant p-value, and the trick suggested in Ref. [48, 52] is to make sure that each permuted p-value is below the permuted p-value for the least significant hypothesis, that

is, $p_{(m-1),i} \leq p_{(m),i}$ for each $i = 1, \ldots, B$. This is the idea behind computation of the successive minima $q_{i,b}$. After this, the second least significant adjusted p-value can be estimated as the proportion of successive minima below the observed $p_{(m-1)}$, and the procedure iterated until the most significant p-value $p_{(1)}$ is used. Finally, monotonicity of the estimated adjusted p-values is enforced to preserve the ordering implied by the observed p-values.

An approach closely connected with the minP procedure is given in Ref. [40] and further studied in Refs. [45, 69], and is usually referred to as max T method. The procedure is a straightforward dual to a min P method, main difference is that it uses only the test statistics. Let $F'_{r,1-\alpha}(\cdot)$ indicate the $1-\alpha$ percentile of the distribution of the maximum of the last r test statistics. Suppose, without loss of generality that the test statistics are ordered. The 'Step-down max T' procedure fixes $C_j = F'^{-1}_{j,1-\alpha}(T_n(m-j+1), \ldots, T_n(m))$, and proceeds in a step-down fashion stopping the first time $T_n(j) \leq C_j$; and rejecting the hypotheses corresponding to $T_n(1), \ldots, T_n(j-1)$.

The functions $F'_{r,\alpha}(\cdot)$ can be estimated through permutation in the usual fashion. Although being based on test statistics, the methods are nevertheless distribution free in FWER control. min P and max T methods are equivalent when the test statistics are identically distributed under the null hypothesis, while may lead to different results otherwise. As a referee pointed out, maxT methods for instance have some advantages in situations in which $n \ll m$.

Genovese and Wasserman[30] propose to use a 'complete null' assumption, thereby letting $F_{j,\alpha}(\cdot)$ be the α-th percentile of a *Beta* distribution with parameters $m-j$ and 1. It is straightforward to see that this choice leads to equivalent step-down constants as in the Step-down Sidak procedure. See also Ref. [69] for further comments on this procedure and an extension under dependence. For the Sidak procedures.[70,71]

A huge amount of work has been done on FWER control. For instance, Finner and Roters[72] compared step-up and step-down procedures, assuming the p-values were exchangeable; showing that step-up procedures controlling the FWER use constants very close to the constants used by step-down procedures, thus being more powerful in general. Dunnet and Tahmane[73] propose a step-up multiple testing procedure, optimal in terms of power, when the test statistics are distributed like a Student's T. Seneta and Chen[74,75] investigate a step-down procedure sharpening step-down Holm, by taking into account the degree of association between the test statistics. We will not attempt to review the great amount of work on FWER here, but instead point the reader for instance.[41,46,76]

4.3 Procedures controlling the FDR

The FDR, as defined in Equation (3), was introduced by Benjamin and Hochberg[27] in response to the need of an error measure that would allow for good power, in particular with large m. While the use of FWER controlling methods is preferable in many situations, they can have low power in certain 'large m' applications. There are no additional assumptions for achieving control of the FDR, apart from considerations on the dependence of p-values (see Section 6). Assumptions on the distribution of p-values under

the alternative or about the true proportion of false nulls may be used to improve the procedures in terms of power, but are usually not needed.

We review four procedures to control the FDR:

BH BH procedure consists in fixing a step-up constant equal to $\alpha_j = j\alpha/m$.[77]

Plug-in A direct improvement of BH procedure is given by using the step-up constant $\alpha_j = j\alpha/m(1-\widehat{a})$, where $\widehat{a}$ is any estimator of a (the proportion of true false hypotheses).[78]

Step-Down BL Step-down BL method consists in fixing the step-down constant $\alpha_j = 1-[1-\min(1,(m/m-j+1)\alpha)]^{1/(m-j+1)}$.[79]

Resampling-based YB[55] suggest to improve power and deal with dependence through the following resampling-based procedure:

1. Bootstrap the data to obtain B vectors of resampled p-values
2. Without loss of generality let the ordered p-values $p_{(k)}$ be the possible thresholds. For each sample let $r(p_{(k)})$ be the number of resampled p-values below $p_{(k)}$, and let $r_\beta(p_{(k)})$ be the $1-\beta$ quantile of $r(p_{(k)})$ for a small β (say $\beta = 0.05$). Then, for each threshold compute $Q^*(p_{(k)})$ as the resample based mean of the function

$$Q(p_{(k)}) = \begin{cases} \dfrac{r(p_{(k)})}{r(p_{(k)}) + k - p_{(k)} * m} & \text{If } p_{(k)} * m \leq k - r_\beta(p_{(k)}) \\ 1 & \text{Otherwise} \end{cases} \tag{9}$$

3. Let $k_\alpha = \max_k\{Q^*(p_{(k)}) \leq \alpha\}$ and set threshold $T = p_{(k_\alpha)}$.

The *BH* procedure was originally proposed in Ref. [77], but it did not receive much attention at that time since it did not control the FWER in the strong sense (while it did in the weak sense). It can be seen that it controls the FDR at level $(1-a)\alpha$, and hence at level α (see,[27] or[80] for a shorter and elegant version of the proof).

Asymptotic results for the BH and plug-in procedures can be found in Ref. [30] and [38]. In particular, Genovese and Wasserman[38] prove that the BH procedure is asymptotically equivalent to a one-step procedure in which the cut-off is the solution u^* of the equation:

$$\frac{F(u)}{u} = \frac{(1-\alpha)(1-a)}{(a\alpha)} \tag{10}$$

where $F(u)$ is the distribution of the p-values under the alternative. If the distributions of the p-values under the alternative are strictly concave and their densities are bounded below at zero by the right hand side of Equation (10), there will be a positive number of rejections. Concavity of $F(\cdot)$ will be given for instance by test statistics whose density is eventually strictly decreasing. Genovese and Wasserman[81] introduce also estimators for a and $F(\cdot)$, suggest ways to build confidence thresholds for the FDP and provide limiting distributions of the quantities of interest. Finally, Sarkar[39] proves that the BH procedure is unbiased, being a special case of FDR-controlling generalized step-up–step-down unbiased procedures proposed in Ref. [82].

The *plug-in* procedure was first proposed in Ref. [78]. This is the only procedure reviewed in this paper that does not achieve strong control of the corresponding error rate. In fact, if all the null hypotheses are true and $a > 0$, it is straightforward to see that FDR control is not guaranteed by the plug-in procedure. The idea is that additional information given by the sequence of p-values can be exploited through a suitable estimator of a. This leads to *exact*, or, in the[78] terminology, *adaptive* control of the FDR, with the advantage of sensibly and often greatly increased power over the BH procedure. It should be furthermore noticed that uncertainty brought about by the estimation of a is not usually incorporated and only asymptotic control may be guaranteed. It is straightforward to see that only if the estimator for a is conservative ($\widehat{a} \leq a$) there is (exact) FDR control and whenever $\widehat{a} > a$ FDR can be above α by a factor of $(1 - \widehat{a})/(1 - a)$. A review of possible estimators for a is given below. A further improvement of the plug-in method can be found in Ref. [83], in which the procedure is repeated iteratively. They note that the two-step version of their procedure can be seen as a first step in which a is estimated, and a second step in which plug-in is applied.

The *step-down BL* procedure was proposed in Ref. [79], who argued with extensive simulations that it neither dominates or is dominated by the BH procedure. In particular, they argue in a large simulation study that it is more powerful than BH procedure when the number of tested hypotheses is small and many of the hypotheses are far from being true. Note that the most common case in applications is that the number of true nulls is the large majority, so that the step-down BL procedure may not be the best choice.

It is worth noticing that resampling-based BY[55] is not guaranteed to yield FDR control. They only argue by simulation that their procedure gives FDR control, and suggest also how to improve it under subset pivotality. Finally, they show an application to the estimation of correlation maps in meteorology.

Storey *et al*.[80] propose a unified estimation approach for the FDR, showing methods to estimate the FDR fixing the threshold or the rejection region, or asymptotically over all rejection regions simultaneously. They suggest a way to control the FDR through their estimates. In particular, they present several theorems that all require almost sure pointwise convergence of the empirical distributions of the subsequence of p-values for which the null is true and the subsequence of p-values for which the alternative hypothesis is true.

Tusher *et al*.[84] introduce the significance analysis for microarrays (SAM), a resampling-based method that controls a functional of the FDP and is appropriately devised for DNA microarray data. As[11] note, it is not clear if SAM does directly control the FDR, even if Storey and Tibshirani[85] suggest possible ways to achieve FDR control. The main problem is that the data are used both to estimate the FDR and the tuning parameters.

Benjamini and Hochberg[36] and Genovese *et al*.[37] describe algorithms that allow to control the propensity of rejection for certain hypotheses if one has prior information. For instance, it is well known that in neurology experiments false nulls are likely to be clustered. Benjamini and Hochberg[36] used the weights in the definition of the error rate (loss weighting), hence changing the error rate to be a 'weighted' false discovery rate. Genovese *et al*.,[37] following,[25] used p-value weighting leaving the error rate unchanged; and showed that if the magnitude of the weights is positively associated with the null

hypothesis being false, power is improved; while power loss is surprisingly negligible for misspecification of the weights.

4.3.1 Estimation of the proportion of false nulls

In this section, we will review some estimators for a, the proportion of false nulls.

The most common estimator used was proposed in Ref. [86], and suggested by Storey[32] for the FDR controlling context. It is defined as:

$$\widehat{a} = \frac{\widehat{G}(t_0) - t_0}{1 - t_0} \tag{11}$$

where $\widehat{G}(t) = 1/m \sum 1_{p_j<t}$ (the empirical distribution of the p-values), and t_0 is fixed in the interval $(0, 1)$. Note that $m\widehat{G}(t_0)$ is simply the count of p-values smaller than t_0. An high t_0 (for instance $t_0 = 0.5$) is used in Ref. [86], since p-values corresponding to true null hypotheses will cluster above high thresholds. A bootstrap method is, instead described in Ref. [32], with the aim of minimizing the (estimated) risk. Another possibility given in Ref. [87] is to choose t_0 as the smallest p-value not rejected by a test for uniformity, for instance a threshold T of any FWER controlling procedure.

Note that Schweder and Spjøtvoll[86] proposed estimator in Equation (11) for an adaptive control of FWER, improving Bonferroni procedure by using the one-step constant $\alpha/m(1 - \widehat{a})$.

Swanepoel,[88] in a different context, proposes a consistent estimator for a defined as $\widehat{a} = \max(0, 1 - \min_{0<x<1} \widehat{g}(x))$; where $\widehat{g}(\cdot)$ is an estimate of the marginal density of the vector of p-values, based on maximal symmetric $2s_n$-spacings. This estimator is seen to converge very fast to the true a but turns out to break down under dependence (simulations not shown).

A slight modification of Equation (11) is in Ref. [89], based on bounding sequences of the weighted empirical distribution of the p-values. The idea is to obtain an estimator that is conservative with high probability. Note that the other estimators reviewed here can lose this conservative property, even under independence.

Another recent result is achieved in Ref. [90], who propose to use a non parametric maximum likelihood estimator of the p-value density yielding an expression very similar to Equation (11): $\widehat{a} = (\widehat{G}(t_0) - t_0)/(p_{(m)} - t_0)$. They also propose a less conservative estimator based on the assumption that the p-value density is decreasing or convex decreasing, and show by simulations that such estimators have got a good performance also under dependence.

Note that estimation of the number of true/false null hypotheses may be of interest per se, especially in applications in functional magnetic resonance imaging[87] or source detection in astrophysics.[88,91]

4.4 Procedures controlling the FDX

FDX, as defined in Equation (4), is a much more recent error rate proposed almost at the same time by van der Laan *et al*.[29] and Genovese and Wasserman.[30]

In FDX control interest is taken in the tails of the distribution of the FDP rather than in its central part. This is useful in cases in which the random variable FDP is not concentrated around its mean, the FDR: while on average it is guaranteed that the proportion of false rejections is low, the realized FDP may be high if there is a weak concentration around the mean. This can be caused for instance by an high variance of the FDP, which can be further increased by dependence among the test statistics. Owen[92] for instance suggests that it may not even be meaningful to control the FDR under dependence since the variance may be too high. FDX control implies that large FDP is realized with small probability; thus being more protective against extremal situations in all cases.

In general, anyway, FDX and FDR control respond differently to the distribution of the p-values under the alternative, $F(\cdot)$; and FDX control may lead to more or less rejections than FDR control on a case by case basis.

We review four procedures to control the FDX:

Augmentation Augmentation was first proposed in Ref. [29]. The steps are given by:

1. Control the FWER with any procedure, and reject $|S_{FWER}|$ hypotheses.
2. If $|S_{FWER}| > 0$, let

$$k_n(c,\alpha) = \max\left\{ j \in \{0,\dots m - |S_{FWER}|\}\colon \frac{j}{j + |S_{FWER}|} \leq c \right\}.$$

   The number $k_n(c,\alpha)$ is easily computed by starting from $j = 0$ and increasing the counter as long as the fraction $j/(j + |S_{FWER}|)$ is below c.
3. Any choice of $k_n(c,\alpha)$ additional hypotheses will control FDX at the desired level. For power considerations, the $k_n(c,\alpha)$ most significant p-values not previously rejected will be selected.

Inversion Inversion was proposed in Ref. [30]. The steps are given by:

1. For every possible subset of p-values test at level α the hypothesis that the p-values are identically distributed like a uniform.
2. Call U the collection of all subsets not rejected in the previous step. The hypotheses in U are candidate to rejection.
3. For any $C \neq \emptyset$ let $\bar{\Gamma}(C) = \max_{B \in U}(|C \cap B|/|C|)$. Let R be the biggest set such that $\bar{\Gamma}(R) \leq c$. R is a rejection set that yields $\mathrm{tFDP}(c) \leq \alpha$.

They also note that the entire procedure can be substituted by a simple augmentation of the Sidak step-down procedure, as we will point out below.

Step-Down LR Lehmann and Romano[33] show that FDX control is achieved by using the step-down constants $\alpha_j = (\lceil cj \rceil + 1)\alpha/(m + \lceil cj \rceil + 1 - j)$.

Resampling-Based LBH A resampling-based procedure is proposed in Ref. [93]:

1. Bootstrap the data, compute the resampled test statistics and center each vector of test statistics by its own mean. This is an estimate of the null distribution $Q_0(t)$.

2. Estimate the density of the test statistics, for instance by bootstrapping. Sample the indicator of each null hypothesis to be false from a Bernoulli with parameter given by an estimated ratio of the null and marginal density $q_0(T_n(j))/g(T_n(j))$.
3. Estimate the realized FDX for each possible cut-off $p_{(1)}, \ldots, p_{(k)}$.
4. Repeat steps 1–3 B times.
5. Estimate the FDX for each cut-off as the average of the realized FDX for each iteration. Set the cut-off for the p-values as the highest cut-off giving FDX below α. More details on this procedure are given below.

Dudoit *et al*.[29] introduce the augmentation procedure. They start from the idea that any procedure requiring something less stringent than FWER control will result in the rejection of at least the same hypotheses. For this reason, they start by controlling the FWER and then they augment by rejecting the previously selected hypotheses and an opportune additional number. In this sense, they propose a universal method to identify additional rejections among the hypotheses which were not rejected with a procedure controlling the FWER (asymptotically or exactly). For power considerations, one obviously adds the k most significant p-values not yet rejected. This method is very flexible and shares the robustness to dependence of FWER controlling procedures. The great advantage of augmentation is in fact that it is valid under general dependence if the FWER is controlled under dependence; and computationally very fast. As we will show later via simulations, the main drawback is that the power may be low for large number of tests. The reason is easily understood just by looking at the definition of $k_n(c, \alpha)$: when no test is rejected at the first stage (something not uncommon for FWER controlling procedures), none will be at the second stage for any $c < 1$.

The resampling-based method in Ref. [93] was proposed mainly to overcome this problem. In their approach, first a null distribution for the test statistics is estimated via the bootstrap or permutation by resampling the data itself and forming each time a vector of test statistics, then each vector is centered by its respective mean. The result is a sample from a null distribution,[51] which we call $Q_0(t)$. Secondly, the indicator of each null hypothesis to be false is sampled from a Bernoulli, with parameter given by an estimated ratio of the null and marginal density $q_0(T_n(j))/g(T_n(j))$. In many applications the null density is known, while the marginal density can be estimated by the bootstrapped test statistics *before* centering. Finally, the realized FDX is estimated and the operation repeated B times. The cut-off for the p-values is set as the highest cut-off giving a proportion of estimated FDX over c smaller than α. The procedure combines a clever null estimation by the bootstrap with an adaptive FDX estimation and control, and it is seen in simulation to be more powerful than the augmentation methods. The main drawback is the high computational cost: in fact it is a double-resampling procedure mostly like the classical resampling Min P of Westfall and Young[40] and this is particularly cumbersome for big m. Moreover, only asymptotic control is guaranteed, and FDX control may not be achieved when the number of observations is very small, like in certain applications in DNA Microarrays data analysis; even if a finite sample rationale is established in Ref. [93]. Code for resampling-based LBH is available for the software R,[94] in the package `multtest`.

The inversion method was proposed in Ref. [30], and involves inverting a set of uniformity tests and forming a confidence upper envelope for the FDP. Genovese and

Wasserman[30] suggest a few possible uniformity tests, among which the min P test of[69] under the complete null (as we pointed out, the step-down Sidak test). They prove that, with this choice, the augmentation and inversion procedures lead to the same rejection regions under mild conditions. For this reason, if step-down Sidak test is used at the first step, the number of uniformity tests reduces from 2^m to m, and this is equivalent to do augmentation of step-down Sidak. We call this '$p_{(1)}$-approach' throughout. Perone Pacifico *et al*.[95] propose a continuous analogue of the inversion method in the context of random fields, where the number of null hypotheses is uncountable; providing also bounds on the (continuous version) of the FNR.

Finally, Lehmann and Romano[33] propose to control the FDX with a step-down procedure (step-down LR) whose constants arise mainly from combinatorial and probabilistic reasoning; and give extensions of their method under dependence which we will discuss in Section 6. We will see in simulations that step-down LR procedure likely dominates the other methods in terms of power.

5 Simulations

We will now provide some simulations in order to illustrate the methods. We point out that the simulation settings are limited, and only narrow generalization of the evidence provided by the simulations in this section is recommended. In particular, we only use independent test statistics, while a broad experiment would require the comparison of different dependence structures. Furthermore, different considerations may arise with the use of different distributions under the alternative.

5.1 Comparison of the multiple testing procedures

In this section, we will briefly compare the procedures on the basis of the counts of errors $N_{1|0}$ and $N_{0|1}$. This provides a direct comparison among procedures in terms of what really happens in applications. We will also use the introduced error measures and the FNR to compare procedures in terms of power. We generated normal random variables, with expected values under the alternative sampled from a random uniform in $(0, 5)$; and then we applied the multiple testing procedures to the vector of p-values arising from one-sided testing with known variance equal to 1.

Table 2 shows the results for $B = 1000$ simulated normal data sets, with $m = 100$ and $M_0 = 90$; Table 3 shows the same for $m = 5000$ and $M_0 = 4500$, and Table 4 for $m = 100\,000$ and $M_0 = 90\,000$. These are realistic situations, as in most applications M_0 is close to m (sparseness).

From the tables we see that the average number of Type I errors tends to grow fast with m under no correction. FWER controlling procedures fail to reject more or less 90% of the true false hypotheses, while FDR controlling methods about a half. This is of course depending on the specific simulation setting, and in particular on the distribution used to sample the normal means under the alternative. Step-down BL procedure performed worse than BH and plug-in (with estimator in Equation (11)) procedures, since $M_0 \cong m$: it is known from Ref. [79] that step-down BL may dominate BH and plug-in in terms of power only when M_0 is far from m. Among the FDX controlling procedures,[33] method

Table 2 Average error counts for $m = 100$ tests, $M_0 = 90$ for different methods controlling different error measures at level $\alpha = 0.05$

Method	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	FDX	FNR
Control of single Type I Error						
Uncorrected	4.49	3.27	0.989	0.386	0.980	0.0366
Control of FWER						
Bonferroni	0.044	6.60	0.044	0.013	0.044	0.068
Step-down Holm	0.048	6.54	0.048	0.015	0.048	0.066
One-step Sidak	0.044	6.59	0.044	0.012	0.044	0.068
Step-down Sidak	0.050	6.48	0.050	0.014	0.050	0.066
Step-up Hochberg	0.048	6.52	0.048	0.014	0.048	0.066
Step-down Min P	0.046	6.46	0.046	0.015	0.046	0.066
Control of FDR						
BH	0.264	5.53	0.221	0.044	0.219	0.058
Plug-in	0.298	5.45	0.241	0.049	0.238	0.057
Step-down BL	0.042	6.44	0.042	0.010	0.042	0.065
BY	0.030	6.657	0.030	0.008	0.030	0.069
Control of tFDP(0.1)						
$p_{(1)}$-approach	0.050	6.48	0.050	0.014	0.050	0.066
Augmentation with Bonferroni at first step	0.044	6.60	0.044	0.013	0.044	0.068
Step-down LR	0.046	6.53	0.045	0.012	0.045	0.067

proves sensibly less stringent than the others. It can be seen that, as m grows, the FDX is slightly lower than its upper bound α. This suggests one could set c as a decreasing function of m, and also that there may be room for improvement of FDX controlling procedures.

Table 3 Average error counts for $m = 5000$ tests, $M_0 = 4500$ for different methods controlling different error measures at level $\alpha = 0.05$

Method	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	FDX	FNR
Control of Single Type I Error						
Uncorrected	225.12	166.12	1.000	0.402	1.000	0.037
Control of FWER						
Bonferroni	0.045	412.68	0.044	0.000	0.000	0.0840
Step-down Holm	0.046	412.36	0.045	0.000	0.000	0.0839
One-step Sidak	0.046	412.23	0.045	0.000	0.000	0.0839
Step-down Sidak	0.047	411.90	0.046	0.000	0.000	0.0838
Step-up Hochberg	0.046	412.36	0.045	0.000	0.000	0.0839
Step-down Min P	0.047	411.90	0.046	0.000	0.000	0.0838
Control of FDR						
BH	10.24	282.40	1.000	0.045	0.000	0.0592
Plug-in	11.51	279.03	1.000	0.049	0.000	0.0585
Step-down BL	0.050	411.59	0.048	0.000	0.000	0.0838
BY	0.735	355.309	0.517	0.005	0.000	0.0732
Control of tFDP(0.1)						
$p_{(1)}$-approach	0.082	402.70	0.079	0.001	0.000	0.0821
Augmentation with Bonferroni at first step	0.077	403.55	0.075	0.001	0.000	0.0822
Step-down LR	0.719	356.74	0.498	0.005	0.000	0.0734

Table 4 Average error counts for $m = 100000$ tests, $M_0 = 90000$ for different methods controlling different error measures at level $\alpha = 0.05$

Method	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	FDX	FNR
Control of single type I error						
Uncorrected	4497.59	3334.47	1.000	0.040	1.000	0.037
Control of FWER						
Bonferroni	0.047	9089.72	0.045	0.000	0.000	0.0917
Step-down Holm	0.047	9087.17	0.045	0.000	0.000	0.0917
One-step Sidak	0.048	9083.76	0.046	0.000	0.000	0.0917
Step-down Sidak	0.048	9081.76	0.046	0.000	0.000	0.0916
Step-up Hochberg	0.047	9087.17	0.046	0.000	0.000	0.0917
Step-down MinP	0.048	9081.77	0.046	0.000	0.000	0.0916
Control of FDR						
BH	203.92	5669.27	1.000	0.045	0.000	0.0594
Plug-in	298.07	5596.54	1.000	0.050	0.001	0.0587
Step-down BL	0.049	9079.70	0.047	0.000	0.000	0.0912
BY	10.15	7291.05	1.000	0.037	0.000	0.0750
Control of tFDP(0.1)						
$p_{(1)}$-approach	0.069	8980.32	0.066	0.000	0.000	0.0907
Augmentation with						
Bonferroni at first step	0.065	8988.43	0.063	0.000	0.000	0.0908
Step-down LR	13.161	7166.89	1.000	0.005	0.000	0.0738

It is the case to make also a comparison across error measures. It can be noted that while the differences within procedures controlling the same error measure are not very marked, there are dramatic differences between error measures. This is well acknowledged. A further remark is that the discrepancies are more and more evident as m grows. It is important to add that there is a clear difference in spirit between FWER and FDR/FDX control; since while FWER is based only on the number of Type I errors, the other two criteria aim at a balance between Type I errors and number of rejections, and hence, power.

Furthermore, while as m changes the controlled error measure is kept more or less constant below α by many procedures, the other error measures can vary wildly; so that simultaneous control of FDR and FDX should not usually be expected.

5.2 FDR control via FDX control: Choice of c

In this section, we will give insights on how to choose a value for c if an FDX controlling procedure is to be used for FDR control. Recall that any FDX controlling procedure controls the FDR at level $c + (1 - c)\alpha$. Hence, if $\text{tFDP}(c) < (\alpha - c)/(1 - c)$, $\text{FDR} \leq \alpha$. We will now get a sense, via simulation, of what happens for different values choices of $c \in (0, \alpha)$. Simulations suggest that the optimal c may usually be close to zero. We will declare the optimal c as the one yielding the lowest FNR as defined in Ref. [5].

Figure 1 shows the results of the simulations for $m = 100$ and $M_0 = 90$. For each value of c, at each of $B = 1000$ iterations we generated normal random variables, with expected values under the alternative sampled from a random uniform in $(0, 5)$; and then we applied the augmentation procedure with Bonferroni at first step to the vector of p-values arising from one-sided testing with known variance equal to 1. The dots

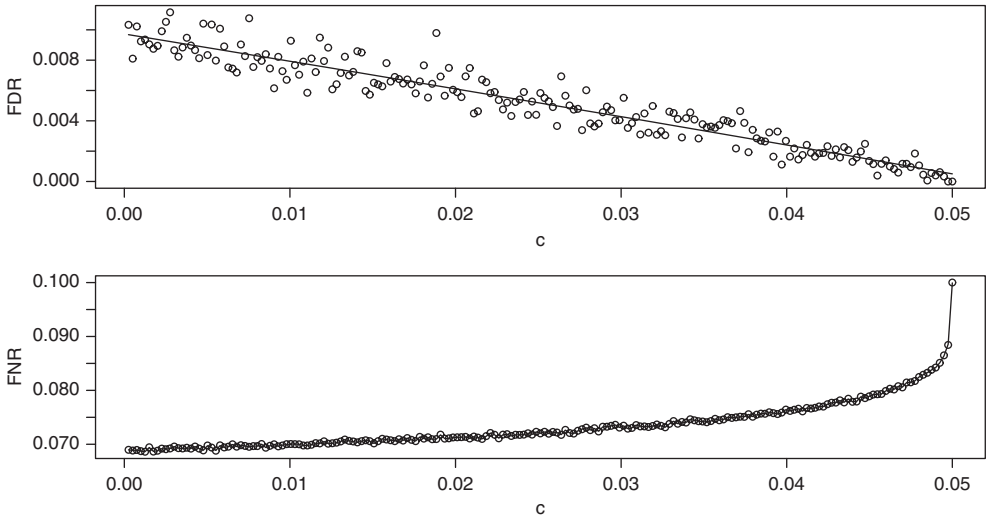

Figure 1 Augmentation procedure, $m = 100$, $M_0 = 90$.

represent the obtained FDR and FNR, while the line is a fitted cubic smoothing spline, with amount of smoothness estimated by cross validation (see for instance[96] for these non-parametric methods). We can see that, unless the number of true nulls is very small, the optimal c is always very close to zero. Moreover, the procedure is always conservative. In this setting, there is a price to pay for using an FDX controlling procedure to control the FDR. Similar results are observed in Figure 2, where simulations are done for $M_0 = 50$. Slightly different conclusions are observed in Figure 3, where $M_0 = 10$, as in that case the low values for the FNR are seen for values of c around 0.035.

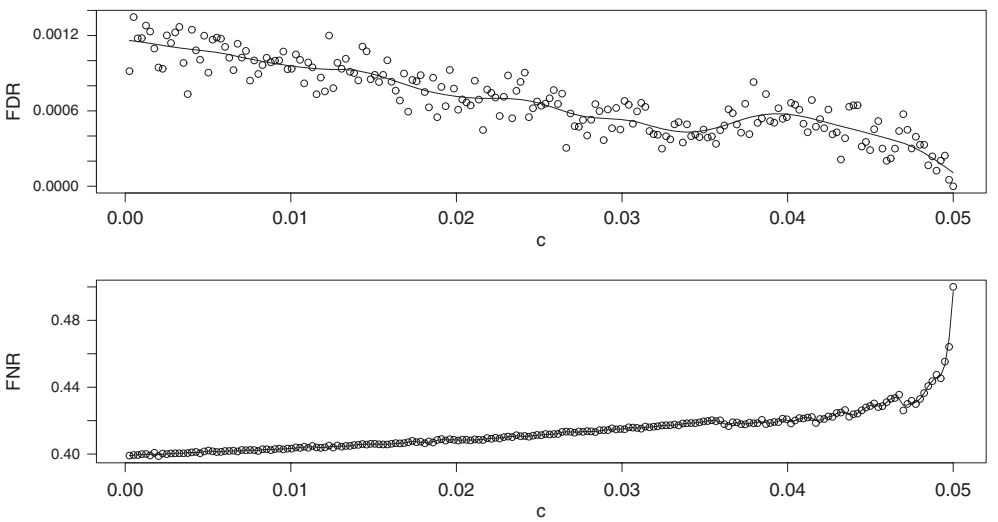

Figure 2 Augmentation procedure, $m = 100$, $M_0 = 50$.

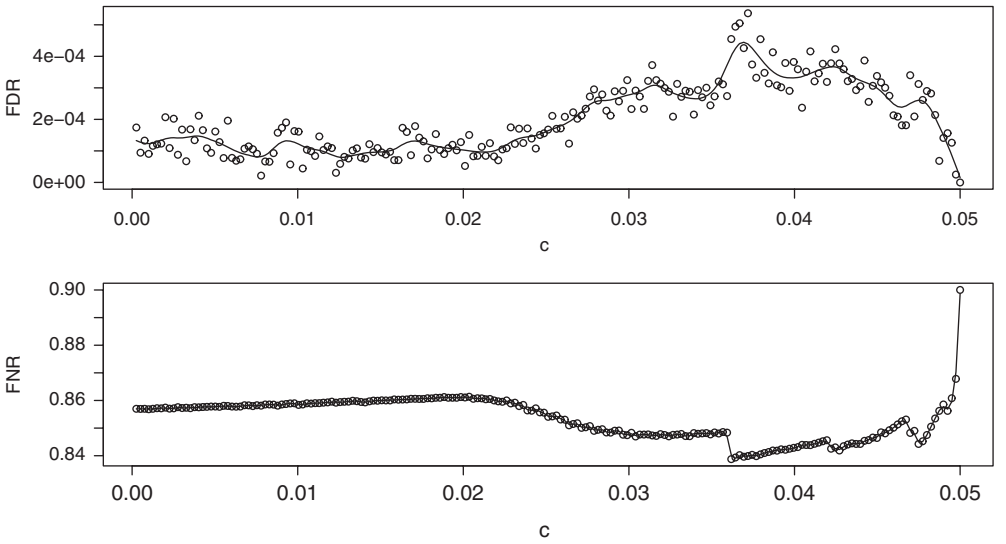

Figure 3 Augmentation procedure, $m = 100$, $M_0 = 10$.

5.3 Bayesian multiple testing: a simulation

As noted by Berry and Hochberg,[97] in the case of multiple testing

> "In the simplest Bayesian view, there is no need for adjustments and the Bayesian perspective is similar to that of the frequentist who makes inferences on a per-comparison basis."

They make a review of available Bayesian procedures to control frequentist error measures, and propose a hierarchical model based on a Dirichlet process prior distribution to allow for exchangeability of the tests. If independent priors are used, they formally conclude that from a Bayesian point of view no modification is needed to the standard single setting procedure. Along the same lines, Berry[98] presented an empirical Bayes approach to multiple testing, utilizing hierarchical models that directly controlled the FWE. This view is strenghtened in Refs. [99] and [100], who claim that a correct adjustment is automatic within the Bayesian paradigm. In this light, Bayesian testing of many hypotheses does not pose different problems than testing a single hypothesis, no adjustment is needed. An early reference for a decision theoretic framework for Bayesian multiple testing is Ref. [101]. Duncan's approach actually corresponds to a different outlook in which there is no true null hypothesis, and the aim is to prevent Type III errors, that is a *directional* error: claiming superiority when there actually is inferiority. Duncan's approach was further studied in this sense by Schaffer.[102] For a discussion on Type III errors see for instance.[103–105]

A way to choose suitable prior distributions on the quantities of interest is proposed in Ref. [99], who also develop efficient importance sampling strategies[106] to deal with multiplicity.

An empirical Bayes framework for FDR controlling procedures is given in Ref. [35] and [107], when proposing the positive false discovery rate, discusses an interesting Bayesian interpretation: the pFDR of the procedure is the posterior probability that a null hypothesis is true given that the corresponding test statistic belongs to a (fixed) rejection region. A Bernoulli prior is assumed for the probability of a null hypothesis being true, and surprisingly this posterior probability does not depend on m. The FDX controlling procedure in Ref. [93] can also be seen to be incorporating a Bayesian prior.

In this section we try to support the view of Berry and Hochberg by simulating a classical Bayesian test and reporting average count of errors and error rates. We will simulate large-m–small-n situations and we will see that the FDR is often automatically controlled. We will test m hypotheses on the mean μ of $n = 30$ normal random variables, with known variance fixed to $\sigma^2 = 1$ and a conjugate normal prior on μ, centered on the null hypothesis and with variance equal to a certain ξ. It is well known that the Bayes factor of the null against the two-sided alternative is then:

$$B(\bar{x}_i) = \sqrt{n\xi + 1} e^{-(\xi n^2 \bar{x}_i^2)/(2(n\xi+1))} \tag{12}$$

where $\bar{x}_i$ is the observed sample average for the i-th test.

We will compare now two choices for the loss function, namely a 0–1 loss, which leads to rejection of all the hypotheses corresponding to a Bayes factor smaller than 1; and a more conservative loss that makes a false positive 19 times worse than a false negative. This leads to rejection in case the posterior probability of the null is smaller than 0.05 (while in the first case the cut-off is 0.5), or in our setting in case the Bayes factor is smaller than 0.0526. For a discussion of generalized 0-1 losses refer to Ref. [108].

We will compare different choices of the prior parameter ξ. The choices $\xi = 1$ and $\xi = 2$ are two default choices, the first being suggested in Ref. [109] to give the prior same weight as that of a single observation, and the second in Ref. [110] to give the prior a shape close to that of a Cauchy with parameters 0 and σ^2, suggested in Ref. [111] as a possible default prior. The other choices of the parameter ξ yield increasingly diffuse priors (clearly, a choice of $\xi = +\infty$ would yield an improper flat prior).

Tables 5 to 7 show the error counts and error measures for respectively $m = 100, 5000, 100\,000$, $n = 30$, $M_0 = 0.9\,\text{m}$ and cut off 1 for the Bayes factor. Tables 8–10 show the error counts and error measures for respectively $m = 100, 5000, 100\,000$, $n = 30$, $M_0 = 0.9\,\text{m}$ and cut-off 0.0526 for the Bayes factor.

Our simulation study is not extensive, but the tables arise anyway the following comments:

- In our setting, there is a strong sensitivity to the prior. Flat priors are in favor of the null hypothesis and tend to make more Type II errors.
- Unlike the uncorrected frequentist setting, the error counts do not explode as the number of tests m grows. In this sense, a correction is not needed, at least for our simple simulation setting.
- If the prior is diffuse enough, both FDR and tFDP(0.1) may be under control.
- FDR and FNR for fixed ξ do not seem to vary very much with m. On the other hand, FDX is strongly affected by changes in the number of tests.

Table 5 Average error counts for $m = 100$ tests, $M_0 = 90$ for Bayesian testing, $n = 30$, cut-off 1

ξ	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	$tFDP(0.1)$	FNR
1	5.43	0.69	0.99	0.35	0.98	0.008
2	3.63	0.75	0.98	0.27	0.91	0.008
15	1.18	0.87	0.70	0.11	0.41	0.009
30	0.82	1.00	0.56	0.08	0.30	0.011
60	0.52	0.99	0.43	0.05	0.18	0.011
75	0.49	1.02	0.38	0.05	0.16	0.011
100	0.41	1.00	0.35	0.04	0.13	0.011
500	0.16	1.10	0.16	0.02	0.06	0.012
1000	0.12	1.18	0.12	0.01	0.05	0.013
1500	0.07	1.24	0.07	0.01	0.03	0.013

Table 6 Average error counts for $m = 5000$ tests, $M_0 = 4500$ for Bayesian testing, $n = 30$, cut-off 1

ξ	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	$tFDP(0.1)$	FNR
1	267.45	34.12	1.00	0.36	1.00	0.008
2	183.72	37.21	1.00	0.28	1.00	0.008
15	60.05	45.50	1.00	0.12	0.90	0.010
30	40.56	47.60	1.00	0.08	0.07	0.010
60	27.64	49.97	1.00	0.06	0.00	0.011
75	24.42	50.91	1.00	0.05	0.00	0.011
100	20.65	51.92	1.00	0.04	0.00	0.011
500	8.50	56.54	1.00	0.02	0.00	0.012
1000	5.79	58.94	1.00	0.01	0.00	0.013
1500	4.80	59.36	0.99	0.01	0.00	0.013

- When we were protective against Type I errors, like in the classical setting, and used a cut off of 0.0526, apart from a single exception FDR and FDX were always controlled at level $\alpha = 0.05$.

As a final comment, note that giving a different weight to each null hypothesis is straightforward in the Bayesian setting.

Table 7 Average error counts for $m = 100\,000$ tests, $M_0 = 90\,000$ for Bayesian testing, $n = 30$, cut-off 1

ξ	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	$tFDP(0.1)$	FNR
1	5361.24	690.57	1.00	0.36	1.00	0.008
2	3679.83	745.10	1.00	0.28	1.00	0.008
15	1199.79	902.68	1.00	0.12	1.00	0.010
30	815.45	951.32	1.00	0.08	0.00	0.010
60	553.69	1001.54	1.00	0.06	0.00	0.011
75	490.71	1014.55	1.00	0.05	0.00	0.011
100	418.11	1031.92	1.00	0.04	0.00	0.011
500	173.96	1132.39	1.00	0.02	0.00	0.012
1000	119.22	1172.65	1.00	0.01	0.00	0.013
1500	94.96	1193.80	1.00	0.01	0.00	0.013

Table 8 Average error counts for $m = 100$ tests, $M_0 = 90$ for Bayesian testing, $n = 30$, cut-off 0.0526

ξ	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	$tFDP(0.1)$	FNR
1	0.18	1.14	0.16	0.017	0.068	0.012
2	0.13	1.14	0.12	0.013	0.040	0.012
15	0.050	1.30	0.06	0.005	0.016	0.014
30	0.04	1.21	0.03	0.004	0.015	0.013
60	0.01	1.36	0.03	0.001	0.005	0.015
75	0.03	1.39	0.02	0.001	0.006	0.015
100	0.02	1.35	0.02	0.002	0.010	0.015
500	0.01	1.47	0.01	0.001	0.004	0.016
1000	0.01	1.49	0.00	0.000	0.003	0.016
1500	0.00	1.48	0.01	0.000	0.002	0.016

Table 9 Average error counts for $m = 5000$ tests, $M_0 = 4500$ for Bayesian testing, $n = 30$, cut-off 0.0526

ξ	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	$tFDP(0.1)$	FNR
1	8.56	56.57	1.00	0.019	0.000	0.012
2	6.47	57.92	1.00	0.014	0.000	0.013
15	2.33	63.55	0.92	0.005	0.000	0.014
30	1.63	65.24	0.82	0.004	0.000	0.014
60	1.19	66.80	0.69	0.003	0.000	0.014
75	1.02	67.65	0.64	0.002	0.000	0.015
100	0.89	68.26	0.59	0.002	0.000	0.015
500	0.39	71.90	0.29	0.001	0.000	0.016
1000	0.25	73.66	0.23	0.001	0.000	0.016
1500	0.19	74.08	0.17	0.000	0.000	0.016

Table 10 Average error counts for $m = 100\,000$ tests, $M_0 = 90\,000$ for Bayesian testing, $n = 30$, cut-off 0.0526

ξ	$E[N_{1\|0}]$	$E[N_{0\|1}]$	FWE	FDR	$tFDP(0.1)$	FNR
1	172.26	1136.39	1.00	0.019	0.000	0.012
2	128.95	1160.87	1.00	0.014	0.000	0.013
15	46.87	1264.93	1.00	0.005	0.000	0.014
30	33.02	1300.85	1.00	0.004	0.000	0.014
60	22.81	1338.02	1.00	0.003	0.000	0.014
75	20.01	1346.49	1.00	0.002	0.000	0.015
100	17.66	1358.77	1.00	0.002	0.000	0.015
500	7.41	1437.57	1.00	0.001	0.000	0.016
1000	5.21	1469.02	0.99	0.000	0.000	0.016
1500	4.13	1486.47	0.98	0.000	0.000	0.016

6 Type I error rates control under dependence

6.1 FWER control

Control of FWER under dependence has never been an issue. It is easily seen that the Bonferroni correction is valid under arbitrary dependence, together with step-down Holm. Step-up Hochberg, like many step-up procedures, is based on so-called

Simes inequality. This inequality can be proved to be valid under certain dependency structures, as proved by Sarkar[112] for positively dependent test statistics. More precisely, Sarkar[112] shows that Simes inequality holds for multivariate totally positive of order two (MTP_2) test statistics. The condition is satisfied by multivariate normals with non-negative correlations; and certain types of multivariate t, F and gamma distributions (see also Ref. [113] for examples of a subclass of MTP_2 random variables). Sidak procedures are valid under the condition of positive orthant dependence: $\Pr(|T_1| \leq t_1, \ldots, |T_m| \leq t_m) \leq \prod \Pr(|T_i| \leq t_i)$; where T_i are the test statistics. Sidak[70] showed that this property holds for instance for any multivariate normal distribution with non-negative correlations, and few other cases, while[114] extended the results to a larger class of distributions, including some t and F distributions. Main reason behind the development of resampling methods is the possibility to efficiently use the information given by dependence.[40] In fact, resampling based step-down minP procedure controls the FWE under arbitrary dependence and may be slightly more powerful than the other methods in certain settings. Dudoit *et al.* [69] propose it for the setting of DNA Microarrays. In that setting, van der Laan and Bryan[115] argue that one needs $n/\log(m) \to \infty$ as $n, m \to \infty$ for consistent estimates of the correlation matrix of the test statistics. Note that in these applications the number of observations n is usually much smaller than the number of p-values m, so that estimators of the null distribution may be unstable in certain cases, as pointed out in Section 3.2.

6.2 FDR and FDX control

When they introduced the FDR, Benjamini and Hochberg[27] proved that the[77] procedure controlled the FDR under independence of the M_0 test statistics corresponding to the true nulls. Providing results under dependence of the whole sequence of p-values has been an open problem since then.

The best results in our opinion are achieved in Ref. [116], who prove that the BH procedure can never control the FDR at level higher than $\alpha \sum_{i=1}^{m} 1/i$ (using an inequality formerly shown in Ref. [117]), so that taking into account a factor of $\sum_{i=1}^{m} 1/i$ will allow to control the FDR under general dependence. We call this corrected and possibly conservative method BY throughout. They also prove that, under conditions of positive regression dependency (PRD) on S_0, the BH procedure is still valid. The condition of PRDS introduced in Ref. [116] is as follows: recall that a set is said to be increasing if for any $x \in D$ and $y \geq x$, $y \in D$. A vector X is PRDS if for any increasing set D and for each $i \in S_0$, $\Pr(X \in D | X_i = x)$ is non decreasing in x. This is a relaxed version of PRD, a slightly more general form of association.[118] Benjamini and Yekutieli[116] note, together with Sarkar,[82] that PRDS distributions include multivariate normal distributions with positive correlations, all uni-dimensional latent variable distributions, and few other cases. Sarkar[82] also extends the results of Ref. [116] by generalizing their results to a whole class of step-up/step-down procedures to control the FDR. An extensive simulation study can be found in Ref. [119].

The plug-in procedure is directly extended under dependence under the same conditions whenever the estimator for the proportion of false nulls a is conservative (i.e.,

$\widehat{a} \leq a$). Another possibility for FDR control under dependence is using the resampling based procedure in Ref. [55], described above.

As said several theorems that prove FDR control of the plug-in procedure under almost sure pointwise convergence of the empirical distributions of the null and alternative *p*-values are given in Ref. [80]. They argue that this may be true also under dependence; and in fact Bickel[120] shows a process with long-range correlations that satisfies the conditions of.[80] It is the case to note that[120] uses the cited example to argue that FDR control under dependence, even if feasible, may not be appropriate and it may be more sensible to control the FDX. Finally, Farcomeni[53] shows that under conditions of weak dependence both plug-in and BH procedures control the FDR, and suggests robust estimators for the proportion of false nulls *a*. The direct consequence is that the BH method and plug-in with a robust estimator for *a* can be used without any correction for the analysis of time-course DNA microarray data, change-point detection in time series, and few other applications.

The *pFDR*, positive false discovery rate of,[35] can be efficiently estimated under dependence for pre-fixed rejection region.[121]

Augmentation procedures that control the FDX are proved by van der Laan *et al*.[29] to be valid under arbitrary dependence, provided the FWER controlling procedure at the first step is valid under dependence. Moreover, the resampling based procedure in Ref. [93] is adaptive and provides asymptotic control also under dependence.

Lehmann and Romano[33] prove their procedure controls the FDX under two forms of dependence. First, they assume independence between the groups of true and false *p*-values, with arbitrary dependence within. Secondly, they adopt the same assumptions of the step-up Hochberg procedure. Finally, they suggest a more conservative procedure controlling the FDX under general dependence (they take into account a factor of $\sum_{i=1}^{\lceil cm \rceil + 1} 1/i$, mimicking[116] results for FDR control).

7 Applications

There is a long list of possible applications of MTPs in the medical literature. We will revisit here a classical example and then focus on two case studies in genomics.

7.1 Multiple endpoints in clinical trials

As said, multiple endpoints analysis is one of the most frequently encountered multiplicity problem in medical research. We will revisit here an example about testing for many endpoints in clinical trials, taken from Ref. [122] and used by Ref. [27] to support the use of FDR.

In a randomized multicentre trial of 421 patients with acute myocardial infarction, a new front-loaded administration of rt-PA (thrombolysis with recombinant tissue-type plasminogen activator) has been compared with APSAC (anisoylated plasminogen streptokinase activator). The treatments are both known to reduce mortality in myocardial infarction.

The researchers defined in this study four families of respectively 11, 8, 6 and 15 hypotheses, without distinction between primary and secondary endpoints:

1. base-line comparisons
2. patency of infarct-related artery
3. reocclusion rates of patent infarct-related artery
4. cardiac and other events after the start of the thrombolitic treatment.

In the fourth family a careful Type I error rate control is desired, since we do not wish to conclude that one treatment is better with respect to a few cardiac or other events if it is merely equivalent to the other one. The ordered p-values from the fourth family are: 0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000; and authors would like to make a statement about reduced in-hospital mortality rate, which corresponds to $p_{(4)}$. If a value of $p_j = 0.0095$ can be declared statistically significant, then rt-PA is to be preferred for reducing the in-hospital mortality rate (and there would also be a difference with respect to other three endpoints). All the FWER controlling procedures considered would lead to the rejection of 3 p-values, together with step-down procedure of Benjamini and Liu, augmentation procedures and step-down LR; thus not supporting a statement about different mortality rate. On the other end, other FDR controlling procedures prove a bit more liberal. The classical procedure of Benjamini and Hochberg leads to rejection of 4 hypotheses, and plug-in to 9, the same number as uncorrected testing; thus supporting a statement about the possibility to decrease in-hospital mortality rate due to myocardial infarction with the use of rt-PA treatment in the clinical course.

7.2 DNA microarrays

7.2.1 The setting of DNA microarrays

We will not attempt here a review of the analysis of DNA microarrays, a field of biostatistics which is receiving more and more attention. We will point the reader to detailed reviews[123–126] also for a survey of the impressive spectrum of biological applications) and only sketch a very simplified explanation of the problem. Refer to Bolsover *et al.*[127] and Garret and Grisham[128] for background on biochemistry and genetics. For a survey on microarray literature, refer also to web sites `http://genomicshome.com` and `http://www.nslij-genetics.org/microarray`, and to `http://www.bioconductor.org` for software support. For a review of multiple testing methods in the context of microarray data analysis see also Ref. [11].

Advances of the technology have made it possible to obtain the expression levels of tens of thousands of genes from a single biological sample. The essential aim of microarray analysis is to measure the messenger RNA abundance in some sampled cells. The result of the experiment is thus a data matrix of n rows (n cDNA samples, in general from the same kind of tissue of individuals in two or more biological conditions) with m columns (one for each gene) with the ($\log_2$) of the expression of the gene. The sample size n typically ranges from 4 to 100 individuals, while the number of genes from a few hundreds to many thousands. It is apparent that a data set in which the number of rows is much smaller than the number of columns presents statistical challenges not traditionally dealt with. Among the immediate statistical issues there is cleaning of such

a massive data set by filtering of bad spots, normalization within and between slides. Finally, a test statistic is computed for each gene (usually, a t or F statistic).

The filtering phase consists in getting rid of badly measured genes. The microarray experiment is subject to a lot of experimental artifacts. For instance, a tiny grain of dust on the slide can invalidate the expression measure for few genes, and the measurements should be discarded. Usually, genes with expression too close to the *saturation* level (the maximum expression recordable by the machine) are not considered in the analysis, together with records not passing a number of quality tests (signal to noise ratio, spot uniformity, dimension of hybridized area, level of background noise, and so on). Then, normalization is performed to get rid of part of the experimental variability and systematic bias inside a single slide and between the slides (for a discussion of the problem, see[129–131]). Again, the conditions under which the experiment is done are crucial to the results: the heat in the room where the experiment is performed, tiny differences in the duration of exposure, and so on. The variability of these conditions introduces bias and increases the variance of the recorded expression levels.

The task of the researcher is then to restrict the number of suspect genes from possibly the whole genome to some tens, which will then be biologically validated with real-time polymerase chain reaction (RT-PCR), RNA blotting, or other techniques: see for instance.[132] Microarrays are often only the first step before further investigation, the *validation* phase. Hence, a small proportion of false positives is allowed; while too many false positives would make the validation phase impossible from an economical point of view: control of FDR or FDX is naturally desirable in the microarray setting, while FWER controlling procedures usually prove to be too strict and end up in an unsatisfactory list of prospective differentially expressed genes.

Even if the problem is sometimes ignored in the literature, it is intuitive that test statistics arising from DNA microarrays are dependent. Genes measured with the same technology in the same laboratory are subject to common sources of noise. Moreover, changes in expression are part of the same biological mechanism, and hence the expression of each gene is not unrelated to the expression of the other genes. Genes are likely to present at least a form of block dependence, with blocks identified by groups of similar mRNA codes and/or more frequently by *pathways*, that is, groups of genes that activate in sequence, structured ordering, or interact. Blocks of dependent genes are reasonably expected to be small (literature investigates pathways of two to five genes, while a maximum of 50 is thought of being possible).

There are many purposes behind microarray experiments. We list four: first, comparing RNA abundance (the expression of a single gene) with the expression of genes from samples of other individuals in different biological conditions, and identify genes that are less expressed (down-regulated) or over expressed (up-regulated) in the biological condition of interest. For instance, if a particular gene is significantly up-regulated in a sample from a group of people having cancer, it is reasonable to view it as associated with the disease. The second purpose can be to identify genes that are *not* expressed. It is well known that a great part of our genome is constituted by DNA that never activates. Again, it is particularly interesting to find genes that that are somehow 'turned off' or 'turned on' by the disease. The third purpose can be identifying the *pathways*. Commonly clustering techniques are applied, like Partitioning Around Medoids (PAM) of Ref. [133]. Clustering is used also to identify groups of active genes without formal

testing. A fourth purpose is to build classifiers, and predict the biological condition. If a good predictor can be formed, then the genes in the classifier are related to the disease; and moreover the disease can be diagnosed by measuring the expression levels of some particular genes. Usually, a test is done on each gene to determine if it is differentially expressed between the biological conditions, and a test is done under each biological condition to determine if the gene is not expressed in that case. The third and fourth task (clustering and classification) are also relevant, in the sense that significance testing is usually performed *before*, in order to select a subset of relevant genes. Using irrelevant genes for class prediction or clustering may lead to inconsistent results. We conclude by saying that in microarray data analysis there usually is no basis for a division of the hypotheses into groups, which must then be considered all together as a single family. This leads to a very large number of tests, and hence to the need of careful and powerful procedures for multiple testing. We will now show two examples of DNA microarray analysis, focusing first on the identification of differentially expressed genes and then on the construction of a classifier.

7.2.2 *Genetic patterns of colon cancer*

Alon *et al*. [134] analyze data on colon cancer. The expression of around 6500 genes is recorded in 40 tumor and 22 normal samples from the colon of 62 patients. After filtering, 2000 genes were normalized, and a two-sample t-statistic was computed for each gene to verify if there was a significant difference between the biological conditions. Figure 4 shows an histogram of the 2000 t-statistics. p-values are computed from the

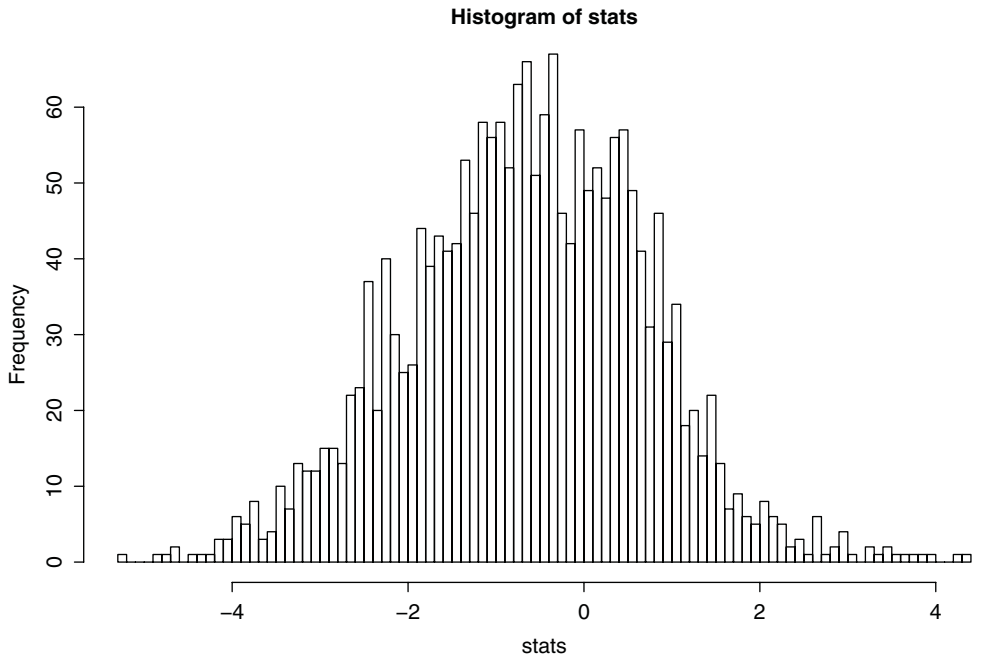

Figure 4 Histogram of 2000 two-sample t-statistics for[134] data.

Table 11 Colon Cancer Data: number of rejections

Bonferroni	11
Step-down Holm	11
One-step Sidak	11
Step-down Sidak	11
Step-up Hochberg	11
Step-down Min P	11
BH	190
BY	38
Plug-in	217
Step-down Benjamini-Liu	11
$p_{(1)}$-approach	12
Augmentation with Bonferroni at first step	12
LR	33

statistics. Colon cancer is well known to be associated with variations in the expression of many genes, and in fact the histogram itself suggests the presence of a few significant genes.

Now, p-values are computed from the test statistics. The number of rejections is summarized in Table 11. It can be seen that the number of rejections here can vary wildly. While in the application of Section 7.1 there just are $m = 15$ tests, here the number of tests is much higher and this leads to a stronger differentiation among FWER, FDR and FDX controlling procedures. BH and plug-in lead to a much higher number of rejections than the other methods.

7.2.3 Classification of lymphoblastic and myeloid leukemia

The approach to cancer classification based on gene expression monitoring by DNA microarrays has been firstly described and applied to human acute leukemia by Golub *et al.*[135] Acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) are two variants of leukemia, which are treated differently. The goal of this experiment is to build a classifier in order to distinguish between the two variants only through gene expression profiling. After pre-processing as described in Ref. [136], 3051 different genes were ready for the analysis, with 27 samples from ALL and 11 from AML.

We decided to build a classifier by using the k-nearest neighbor,[137] with $k = 3$; and by feeding the classifier with the genes selected by a carefully chosen multiple testing procedure.

In order to evaluate the performance of the classifier we split the data into a training set of 15 samples, 9 chosen at random from the ALL samples and 6 chosen at random from the AML samples. A t-test was performed on each gene, obtaining a vector of $m = 3051$ p-values; and multiple testing procedures in Table 12 were used for gene selection. For each set of selected genes the remaining test set of 34 samples, of which 18 ALL and 5 AML, was used to estimate the classification error, that is, the proportion of samples in the test set that were misclassified by the k-NN classifier using the genes selected by the corresponding multiple testing procedure. The operation was repeated 1000 times. The average number of selected genes and the estimated classification error

Table 12 Leukemia Data: average number of rejections and estimated classification error

Bonferroni	15.68	0.0684
Step-down Holm	15.73	0.0683
One-step Sidak	15.88	0.0673
Step-down Sidak	15.93	0.0669
Step-up Hochberg	15.73	0.0682
Step-down Min P	15.92	0.0669
BH	240.46	0.0251
BY	33.66	0.0571
Plug-in	368.45	0.0248
Step-down Benjamini-Liu	15.98	0.0668
$p_{(1)}$-approach	17.11	0.0663
Augmentation with Bonferroni at first step	16.84	0.0666
LR	35.18	0.0521

are reported in Table 12. The results give a further confirmation of the fact that FDR and FDX controlling procedures may be preferred over FWER control in the setting of DNA microarray analysis. In this application in fact they achieve a good balance between number of genes used by the classifier, and classification error (and consequently relevance of the selected genes).

8 Discussion

Multiple hypothesis testing is concerned with maintaining low the number of false positives when testing several hypotheses simultaneously, while retaining a reasonable power for tests of the individual hypotheses. A multiple testing situation presents many substantial differences with the single hypothesis setting, and in particular, a correction on the significance level is needed.

There are quite a few distinct and competing methodologies to deal with multiple tests, some of which we reviewed and we summarize in Table 13. Unless stated otherwise, strong control of the error measure is achieved. We did not attempt to propose a unification here, but rather argued that this diversity is a tool to the researcher, who should know the properties and behavior of the procedures in the different situations that go under the wide multiple hypotheses framework. No one solution is acceptable for all situations.

In this review we tried to argue that when doing many tests at once 1) a sensible choice of Type I error rate is to be made, 2) control of this error rate is to be achieved through a correction, that is, a multiple testing procedure; and 3) the multiple testing procedure is to be chosen as the possibly most powerful for the experimental situation (dependence, possibility to do resampling, information about the proportion of false nulls, and so on). Research in the area of multiple testing is open from the point of view of giving guidelines to choose the Type I error rate, and an optimality criterion, devising powerful and robust procedures especially under dependence of the test statistics, and finally promulgating FDP controlling procedures in certain applications.

Table 13 Multiple testing procedures and their characteristics

Name	Type	Control	Dependence
Control of single Type I error			
Uncorrected	One-step	Finite sample	Arbitrary
Control of FWER			
Bonferroni	One-step	Finite sample	Arbitrary
Step-down Holm	Step-down	Finite sample	Arbitrary
One-step Sidak	One-step	Finite sample	Positive orthant
Step-down Sidak	Step-down	Finite sample	Positive orthant
Step-up Hochberg	Step-up	Finite sample	MTP_2
Step-down Min P	Step-down (Permutation)	Finite sample	Arbitrary
Control of FDR			
BH	Step-up	Finite sample	PRDS
Plug-in	Step-up	(*Exact*) Finite sample/Asymptotic	PRDS
BL	Step-down	Finite sample	Independence
YB	Step-up (Bootstrap)	Asymptotic (?)	Arbitrary
BY	Step-up	Finite sample	Arbitrary
Control of FDX			
$p_{(1)}$-approach	Augmentation/step-down	Finite sample	Positive orthant
Augmentation with Bonferroni at first step	Augmentation	Finite sample	Arbitrary
LR	Step-down	Finite sample	Positive
Lehmann and Romano conservative version	Step-down	Finite sample	Arbitrary
van der Laan Birkner and Hubbard	Bootstrap	Asymptotic	Arbitrary

While there practically was the only choice of the FWE up to some years ago; nowadays there is a continuously growing number of Type I error rates among which to choose. We stressed that these error rates are not only different in the number of rejected hypotheses. While methods that control the FWE (and generalizations) rely only on the number of Type I errors, control of other error rates (FDR, FDX, and generalizations) aims at a balance between Type I and Type II errors.

FDR/FDX control is nowadays a common choice in certain applications mainly because of two reasons: in these applications m is too large to control stringent error rates like the FWE, and moreover statistical procedures serve often only as an exploratory tool for pre-screening of the hypotheses, and after that a formal confirmation takes place. For instance in DNA Microarray studies among the genes declared significant certain are selected for more accurate validation using low-throughput procedures like polymerase chain reaction.

The most pressing advance is in our opinion the discovery of a multiple testing procedure that controls the FDR under arbitrary dependence but it is competitive with BH method in terms of power. As we saw, moreover, there may be room for improvement in terms of power for FDX controlling procedures, and the 'default' choice of $c = 0.1$ should probably be discussed more formally in dependence of the specific application and the number of tests at stake.

There are many other open questions for research in multiple testing: until now for instance the literature on FDP and multiple testing in general does not seem to be

interested in extensions to composite null hypotheses. It is well known that when the null hypothesis is composite the interpretation of p-values is more complex (see for instance[138]), and furthermore the distribution under the null hypothesis need not be uniform. The only practical solution at this point seems to be estimation of p-values through resampling. As an aside, we refer the reader to,[139] where an objective Bayesian approach is used to derive alternative significance levels (that is, alternative p-values) that are uniformly distributed under the null. If one makes use of their *partial posterior predictive p-value*, *U-conditional predictive p-value*, or similar alternative p-values; the procedures may be directly extended to the case of composite null hypotheses. A similar idea is developed in Ref. [140]. Among other open problems, there is the derivation of a framework for power analysis; and a closely related problem, that is, a method to choose the sample size for each test. This has been tackled for instance by Müller *et al*.[141] in the setting of DNA Microarrays.

Acknowledgements

The author is grateful to the editor and four referees for detailed and careful comments that enhanced the completeness and clarity of the paper; and to Livio Finos for advice.

References

1 Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* 1996; **4**: 58–73.
2 Ellis SP, Underwood MD, Arango V. Mixed models and multiple comparisons in analysis of human neurochemical maps. *Psychiat Res-Neuroim* 2000; **9**: 111–19.
3 Merriam EP, Genovese CR, Colby CL. Spatial updating in human parietal cortex. *Neuron* 2003; **39**: 361–73.
4 Logan BR, Rowe DB. An evaluation of thresholding techniques in fMRI analysis. *Neuroimage* 2004; **22**: 95–108.
5 Drigalenko EI, Elston RC. False discoveries in genome scanning. *Genetics Epidemiology* 1997; **14**: 779–84.
6 Weller JI, Song JZ, Heyen DW. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* 1998; **150**(4): 1699–706.
7 Heyen DW, Weller JI, Ron M. A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiological Genomics* 1999; **1**(3): 165–75.
8 Bovenhuis H, Spelman RJ. Selective genotyping to detect quantitative trait loci for multiple traits in outbred populations. *Journal of Dairy Science* 2000; **83**(1): 173–80.
9 Mosig MO, Lipkin E, Khutoreskaya G, Tchourzyna E, Soller M, Fridmann AA. Whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* 2001; **157**: 1683–98.
10 Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 2003; **19**: 368–75.
11 Dudoit S, Shaffer PJ, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science* 2003; **18**: 71–103.
12 Sebastiani P, Gussoni E, Kohane IS, Ramoni MF. Statistical challenges in functional genomics. *Statistical Science* 2003; **18**: 33–70.
13 Khatri P, Babyyak M, Croughwell ND. Temperature during coronary artery bypass

surgery affects quality of life. *Annals of Thoracical Surgery* 2001; **71**(1): 110–16.
14 Schlaeppi M, Edwards K, Fuller RW. Patient perception of the Diskus inhaler: a comparison with the Turbuhaler inhaler. *British Journal of Clinical Practice* 1996; **50**: 14–19.
15 Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology* 1998; **147**: 615–19.
16 Vedantham K, Brunet A, Boyer R. Post-traumatic stress disorder, trauma exposure, and the current health of Canadian bus drivers. *Canadian Journal of Psychiatrics* 2001; **46**(2): 149–55.
17 Schaffer CM, Green PE. Cluster-based market segmentation: some further comparisons of alternative approaches. *Journal of Market Research Sociology* 1998; **40**: 155–63.
18 George EI. The variable selection problem. *Journal of the American Statistical Association* 2000; **95**(452): 1304–08.
19 George EI, Foster DP. Calibration and empirical Bayes variable selection. *Biometrika* 2000; **87**(4): 731–47.
20 Ip EH. Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika* 2001; **66**(1): 109–32.
21 Green SB, Babyak MA. Control of type I errors with multiple tests constraints in structural equation modeling. *Multivariable Behavioural Research* 1997; **32**: 39–51.
22 Yekutieli D, Reiner-Benaim A, Benjamini Y, Elmer GI, Kafkafi N, Letwin NE, Lee NH. Approaches to multiplicity issues in complex research in microarray analysis. *Statistica Neerlandica* 2006; **60**: 414–37.
23 Abramovich F, Benjamini Y. Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis* 1996; **22**: 351–61.
24 Abramovich F, Benjamini Y, Donoho D, Johnstone I. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics* 2006; **34**: 584–653.
25 Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**: 65–70.
26 Westfall PH, Kropf S, Finos L. Weighted FWE-controlling methods in high-dimensional situations. In Benjamini Y, Bretz F, Sarkar S, eds. *Recent developments in multiple comparison procedures*. vol. 47. Institute of Mathematical Statistics Lecture Notes–Monograph Series, 2004: 143–54.
27 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society (Ser B)* 1995; **57**: 289–300.
28 Seeger P. A note on a method for the analysis of significance *en masse*. *Technometrics* 1968; **10**: 586–93.
29 van der Laan MJ, Dudoit S, Pollard KS. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1).
30 Genovese CR, Wasserman L. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association* 2006; **101**: 1408–17.
31 Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetics Epidemiology* 2002; **23**: 70–86.
32 Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society (Ser B)* 2002; **64**: 479–98.
33 Lehmann EL, Romano JP. Generalizations of the familywise error rate. *Annals of Statistics* 2005; **33**: 1138–54.
34 Sarkar SK. *Stepup procedures controlling generalized FWER and generalized FDR*. Department of Statistics, Temple University, 2005.
35 Storey JD. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Annals of Statistics* 2003; **31**: 2013–35.
36 Benjamini Y, Hochberg Y. Multiple Hypothesis Testing with weights. *Scandinavian Journal of Statistics* 1997; **24**: 407–18.
37 Genovese CR, Roeder K, Wasserman L. False discovery control with *p*-value weighting. *Biometrika* 2006; **93**: 509–24.
38 Genovese CR, Wasserman L. Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society (Ser B)* 2002; **64**: 499–518.
39 Sarkar SK. FDR-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference* 2004; **125**: 119–37.

40 Westfall PH, Young SS. *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, 1993.
41 Miller RG. *Simultaneous statistical inference*. Wiley, 1981.
42 Diaconis P. Theories of data analysis: from magical thinking through classical statistics. In Hoaglin DC, Mosteller F, Tukey JW, eds. *Exploring Data Tables, Trends, and Shapes*. Wiley, 1985.
43 Ahmed SW. Issues arising in the application of Bonferroni procedures in federal surveys. In *1991 ASA Proceedings of the Survey Research Methods Section*, 1991. 344–49.
44 Wright SP. Adjusted *p*-values for simultane ous inference. *Biometrics* 1992; **48**: 1005–10.
45 Dudoit S, van der Laan MJ, Pollard KS. Multiple Testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1).
46 Hochberg Y, Tamhane AC. *Multiple comparisons procedures*. Wiley, 1987.
47 Finner H, Roters M. Multiple hypotheses testing and expected number of Type I errors. *Annals of Statistics* 2002; **30**: 220–38.
48 Pesarin F. *Multivariate permutation tests with applications to biostatistics*. Wiley, 2001.
49 Efron B, Tibshirani R. *An introduction to the Bootstrap*. Springer-Verlag, 1993.
50 Troendle K, McShane. An example of slow convergence of the Bootstrap in high dimensions. *American Statistician* 2004; **58**: 25–9.
51 Pollard KS, van der Laan MJ. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* 2004; **125**: 85–100.
52 Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Test* 2003; **12**: 1–77.
53 Farcomeni A. Some Results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics* 2006; published ahead of print. doi: 10.1111/j.1467-9469.2006.00530.x
54 Ferreira JA, Zwinderman AH. On the Benjamini-Hochberg method. *Annals of Statistics* 2006; **34**: 1827–49.
55 Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 1999; **82**: 171–96.
56 Meinert CL. *Clinical trials design, conduct, and analysis*. Oxford University Press, 1986.
57 Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Controlled Clinical Trials* 1997; **18**: 530–45.
58 Chi GYH. Multiple testings: multiple comparisons and multiple endpoints. *Drug Information Journal* 1998; **32**: 1347S–62S.
59 Moyé LA. *P*-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology* 1998; **8**: 351–57.
60 Moyé LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in Medicine* 2000; **19**: 767–79.
61 O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**: 1079–87. With errata: *Biometrics* 1995; **51**: 1580–81.
62 Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**: 487–98.
63 Follmann D. Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine* 1995; **14**: 1163–75.
64 Läuter J. Exact *t* and *F* tests for analyzing studies with multiple endpoints. *Biometrics* 1996; **52**: 964–70.
65 Wei LJ, Lachin JM. Two-sample asymptotially distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* 1984; **79**: 653–61.
66 Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**: 1065–73.
67 Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* 1991; **47**: 511–21.
68 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75** :800–2.
69 van der Laan MJ, Dudoit S, Pollard KS. Multiple Testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1).
70 Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American*

Statistical Association 1967; **62**: 626–33.
71 Sidak Z. On probabilities of rectangles in multivariate Student distributions: their dependence on correlations. *Annals of Mathematical Statistics* 1971; **42**: 169–75.
72 Finner H, Roters M. Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. *Annals of Statistics* 1998; **26**: 505–24.
73 Dunnet CW, Tamhane AC. A Step-Up Multiple Test Procedure. *Journal of the American Statistical Association* 1992; **87**: 162–70.
74 Seneta E, Chen JT. A sequentially rejective test procedure. *Theory of Stochastic Processes* 1997; **3**: 393–402.
75 Seneta E, Chen JT. Simple stepwise tests of hypotheses and multiple comparisons. *International Statistical Review* 2005; **73**: 21–34.
76 Shaffer J. Multiple hypothesis testing. *Annals of Review Psychology* 1995; **46**: 561–84.
77 Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**: 751–54.
78 Benjamini Y, Hochberg Y. The adaptive control of the false discovery rate in multiple hypothesis testing with independent test statistics. *Journal of Educational Behaviour Statistics* 2000; **25**: 60–83.
79 Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 1999; **82**: 163–70.
80 Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society (Ser B)* 2004; **66**: 187–205.
81 Genovese CR, Wasserman L. A stochastic process approach to false discovery control. *Annals of Statistics* 2004; **32**: 1035–61.
82 Sarkar SK. Some Results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics* 2002; **30**: 239–57.
83 Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step up procedures that control the false discovery rate. *Biometrika* 2006; **93**: 491–507.
84 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 2001; **98**: 5116–21.
85 Storey JD, Tibshirani R. *Estimating false discovery rates under dependence, with applications to DNA microarrays*. Department of Statistics, Stanford University, 2001.
86 Schweder T, Spjøtvoll E. Plots of *p*-values to evaluate many hypotheses simultaneously. *Biometrika* 1982; **69**: 493–502.
87 Turkheimer FE, Smith CB, Schmidt K. Estimation of the number of 'true' null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage* 2001; **13**: 920–30.
88 Swanepoel JWH. The limiting behavior of a modified maximal symmetric 2*s*-spacing with applications. *Annals of Statistics* 1999; **27**: 24–35.
89 Meinshausen N, Rice J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics* 2006; 34: 373–393.
90 Langass M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data.*Journal of the Royal Statistical Society (Ser B)* 2005; **67**: 555–72.
91 Miller CJ, Genovese C, Nichol RC, Wasserman L, Connolly A, Reichart D, Hopkins A, Schneider J, Moore A. Controlling the false-discovery rate in astrophysical data analysis. *Astronological Journal* 2001; **122**: 3492–505.
92 Owen AB. Variance of the number of false discoveries. *Journal of the Royal Statistical Society (Ser B)* 2005; **67**: 411–26.
93 van der Laan MJ, Birkner MD, Hubbard AE. Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* 2005; **4**(1).
94 R Development Core Team. R. *A language and environment for statistical computing*. R Foundation for Statistical Computing, 2004.
95 Perone Pacifico M, Genovese C, Verdinelli I, Wasserman L. False discovery control for random fields. *Journal of the American Statistical Association* 2004; **99**: 1002–14.

96 Green PJ, Silverman BW. *Nonparametric regression and generalized linear models: A Roughness Penalty Approach*. Chapman & Hall, 1994.

97 Berry DA, Hochberg Y. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference* 1999; **82**: 215–77.

98 Berry DA. Multiple comparisons, multiple tests, and data dredging: a Bayesian perspective. In Bernardo J, DeGroot M, Lindley D, Smith A, eds. *Bayesian statistics*. vol. 3. Oxford University Press, 1988: 79–94.

99 Scott JG, Berger JO. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* 2006; **136**: 2144–62.

100 Bayarri MJ, Berger JO. The interplay of Bayesian and frequentist analysis. *Statistical Science* 2004; **19**: 58–80.

101 Duncan DB. A Bayesian approach to multiple comparisons. *Technomerics* 1965; **7**: 171–222.

102 Shaffer JP. A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *Journal of Statistical Planning and Inference* 1999; **82**: 197–213.

103 Leibermann B. *Contemporary problems in statistics*. Oxford University Press, 1971.

104 Finner H. Stepwise multiple test procedures and control of directional errors. *Annals of Statistics* 1999; **27**: 274–89.

105 Shaffer JP. Multiplicity, directional (Type III) errors, and the null hypothesis. *Psychological Methods* 2002; **7**: 356–69.

106 Robert CP, Casella G. *Monte Carlo statistical methods*. Springer-Verlag, 1999.

107 Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**: 1151–60.

108 Bernardo JM, Smith AFM. *Bayesian theory*. Wiley and Sons, 1994.

109 Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. *Journal of the American Statistical Association* 1995; **90**(431): 928–39.

110 Berger JO, Boukai B, Wang Y. Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science* 1997; **12**(3): 133–60.

111 Jeffreys H. *Theory of probability*. Oxford University Press, 1961.

112 Sarkar SK. Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Annals of Statistics* 1998; **26**: 494–504.

113 Sarkar SK, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**: 1601–8.

114 Jogdeo K. Association and probability inequalities. *Annals of Statistics* 1977; **5**: 495–504.

115 van der Laan MJ, Bryan J. Gene expression analysis with the parametric Bootstrap. *Biostatistics* 2000; **1**: 1–19.

116 Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001; **29**: 1165–88.

117 Hommel G. Tests of the overall hypothesis for arbitrary dependence structures. *Biomedical Journal* 1983; **25**: 423–30.

118 Esary JD, Proschan F, Walkup DW. Association of random variables, with applications. *Annals of Mathematical Statistics* 1967; **38**: 1466–74.

119 Farcomeni A. More powerful control of the false discovery rate under dependence. *Statistical Methods & Applications* 2006; **15**: 43–73.

120 Bickel DR. *On 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates': does a large number of tests obviate confidence intervals of the FDR?* Office of Biostatistics and Bioinformatics, Medical College of Georgia, 2004.

121 Storey JD, Tibshirani R. Statistical significance for genome-wide studies. In *Proceedings of the National Academy of Sciences* 100, 2003: 9440–5.

122 Neuhaus KL, Von Essen R, Tebbe U, Vogt A, Roth M, Riess M, Niederer W, Forycki F, Wirtzfeld A, Mauurer W, Limbourg P, Merx W, Haerten K. Improved thrombolysis in acute myocaridial infarction front-loaded administration of Alteplase: results of the rt-PA-APSAC patency study (TAPS). *JAmCollCard* 1992; **19**: 885–91.

123 Amaratunga D, Cabrera J. *Exploration and analysis of DNA microarray and protein array data*. Wiley, 2004.

124 Parmigiani G, Garret ES, Irizarry R, Zeger SL. *The analysis of gene expression data: methods and software*. Springer, 2003.

125 Brown PO, Botstein D. Exporing the new world of genome with DNA microarrays. *Nature Genetics* 1999; **21**: 33–7.
126 Duggan D, Bittner M, Chen Y, Meltzer P, Trent J. Expression profiling usign cDNA microarrays. *Nature Genetics* 1999; **21**: 10–14.
127 Bolsover SR, Hyams J, Jones S, Shepard EA, White HA. *From genes to cells*. Wiley, 1997.
128 Garret RH, Grisham CM. *Principles of byochemistry*. Brooks/Cole, 2002.
129 Tseng G, Oh M, Rohlin L, Liao J, Wong W. Issues on cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* 2001; **29**: 2549–57.
130 Yang YH, Dudoit S, Luu P, Speed TP. Normalization for cDNA microarray data. *SPIE BiOS* 2001; 2001.
131 Durbin BP, Rocke DM. Variance stabilizing transformations for two-color microarrays. *Bionformatics* 2004; **20**: 660–7.
132 Zweiger G. *Transducing the genome: information, anarchy and revolution in the biomedical sciences*. McGraw-Hill, 2001.
133 Kaufman L, Rousseeuw PJ. *Finding groups in data*. Wiley, 1990.
134 Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. *Proceedings National Academic Science in the USA* 1999; **96**: 6745–50.
135 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov HJP andColler. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–7.
136 Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 2002; **97**: 77–87.
137 Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 1967; **IT-13**: 21–7.
138 Schervish M. *p*-values: What they are and What they are not? *American Statistician* 1996; **50**: 203–6.
139 Bayarri MJ, Berger JO. *p*-values for composite null models. *Journal of the American Statistical Association* 2000; **95**: 1127–42.
140 Cabras S. *Control of the false discovery rate with frequentist p-values in microarray data analysis*. Università degli studi di Firenze, 2004.
141 Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* 2004; **99**: 990–1001.