Discoveries at Risk

Nicolai Meinshausen and Peter Bühlmann
Seminar für Statistik, ETH Zürich, Switzerland

May 9, 2003

Abstract

When testing multiple hypotheses simultaneously, the false discovery rate (*FDR*) measures the expected proportion of falsely rejected hypothesis. The true amount of false discoveries, however, is very often much larger than indicated by *FDR*. We propose the new Discoveries-at-Risk approach (*DaR*) to multiple hypotheses testing, a generalization of the family-wise error rate (*FWER*). *FWER* can still be controlled, if desired, but more powerful testing is possible by allowing a certain fraction of false discoveries. This is in common with the *FDR*-approach to multiple hypotheses testing. The risk of underestimating the true proportion of false discoveries is, however, tightly controlled in the Discoveries-at-Risk approach.

Although *DaR* often pays a price in terms of power for such tighter control of underestimation compared to *FDR*, we present a surprising result saying that our new *DaR* approach offers both tighter control and more power than *FDR* when controlling at low error rates.

The proposed method of *DaR*-control is applied to simulated and microarray data and compared to *FWER*- and *FDR*- controlling procedures.

1 Introduction

With growing amount of available data in diverse fields, notably biology, simultaneous testing of multiple hypotheses has become increasingly popular. It is misleading to test each of possibly thousands of hypotheses individually and declare those with a sufficiently low p-value as "significant findings". A large proportion of these "discoveries" might be due to falsely rejected null hypotheses, see e.g. Soric (1989).
There has been growing interest recently to develop appropriate type I error rates in multiple testing situations. The two most prominent examples are the family-wise error rate, see e.g. Holm (1979) or Westfall and Young (1993), and the false discovery rate, introduced by Benjamini and Hochberg (1995).

1.1 Preliminaries

In a general multiple testing situation, we assume that there are m hypotheses to test, m_0 of which are null hypotheses and $m_1 = m - m_0$ that fulfill the alternative hypothesis.
Each hypothesis is rejected if the p-value of the corresponding test is in the rejection region $\Gamma = [0, \gamma]$. The total number of rejected hypotheses is a random variable and denoted by

	accepted	rejected	
null true	U	V	m_0
alternative true	T	S	m_1
	W	R	m

Table 1: Notation. Number of correctly and falsely accepted and rejected hypotheses.

R. The realization of this random variable is the observed number of rejections R_{obs}. The total number of rejected null hypotheses is denoted by V. Table 1 below summarizes the notation. Whenever necessary, dependence upon the rejection region $\Gamma = [0, \gamma]$ is indicated by $V(\gamma)$, $R(\gamma)$ etc.

1.2 Type I error rates

The family-wise error rate ($FWER$) is defined as the probability of making at least one wrong rejection:

$$FWER = P[V > 0].$$

The $FWER$ measures thus the risk that one or more null hypotheses are falsely rejected. Control of $FWER$ is often criticized for being too conservative, resulting in low power. In fact, the power vanishes for $m_0 \to \infty$ under fairly general assumptions as shown in section 2.

Often one does not care about some false rejections as long as they are rare compared to the total number of rejections. An interesting quantity is thus the proportion of falsely rejected null hypotheses among all rejections. The false discovery rate (FDR), proposed by Benjamini and Hochberg (1995), is the expectation of this ratio. It is defined as

$$FDR = E[Q] \qquad \text{with } Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} \tag{1}$$

False discoveries are "permitted" as long as they are rare compared to the number of correctly rejected hypotheses.

It has to be noted, though, that FDR measures only the *expected* proportion of falsely rejected hypotheses. FDR contains no information about the variance or distribution of this quantity. The variance can be quite high, in particular for dependent test statistics. A serious, yet mostly ignored shortcoming of FDR is the high risk that the actual proportion of falsely rejected hypotheses is much larger than suggested by FDR.

1.3 Outline

In section 2 the new and flexible framework of "Discoveries-at-Risk" is introduced as a generalization of the family-wise error rate. $FWER$ can still be controlled in the new approach, if desired, but more powerful testing is possible by allowing a certain fraction of false discoveries. This is in common with the FDR-approach to multiple hypotheses testing. The risk of underestimating the true proportion of false discoveries is, however, tightly controlled in the Discoveries-at-Risk approach.

The relations to the family-wise error rate and the false discovery rate are shown in sections 2.3 and 2.4. Section 2 contains most of the theoretical arguments why *DaR* should often be preferred over *FWER* and *FDR*. In particular, we present a surprising result saying that in comparison to *FDR*, our new *DaR* approach offers both tighter control of underestimating the actual amount of false discoveries as well as more power when controlling at low error rates.
A possible estimate of *DaR*, which is constructive and bounds *DaR* from above, is presented in section 3.
Last, we show with simulated data in section 4 that *FDR* indeed frequently underestimates the true proportion of false discoveries. We then apply Discoveries-at-Risk control to the problem of detecting differentially expressed genes from publicly available microarray data and compare the results to various *FWER*- and *FDR*- controlling procedures.

2 Discoveries-at-Risk

2.1 Discoveries-at-Risk for fixed Rejection Regions

In a typical multiple testing situation we can approximate or at least bound from above the distribution of falsely rejected hypotheses V. To assess the significance of the rejections made, one could calculate the probability that the number of false rejections is at least as large as the actually observed number of rejections,

$$P[V \geq R_{obs}].$$

If this probability is substantial, one cannot even exclude the possibility that all rejections are false. In most multiple testing problems, however, this probability is very small. In other words, it is very unlikely that false rejections account for *all* rejections made.
On the other hand, it is often very likely that *some* of the true null hypotheses are falsely rejected. In other words, the value of

$$P[V > 0],$$

or, equivalently, the value of the family-wise error rate is often large.
As an intermediate approach we are interested in finding the smallest proportion β of all rejections R_{obs}, such that

$$P[V > \beta R_{obs}] \tag{2}$$

is bounded by some small α. It is then very unlikely that false rejections account for more than this fraction of all rejections made.

Definition 1 (Discoveries-at-Risk) *Let $q_{1-\alpha} = q_{1-\alpha}(\gamma)$ be the (1-α)-quantile of the distribution of $V(\gamma)$. Discoveries-at-Risk at level α is defined as*

$$DaR^{\alpha}(\gamma) = \begin{cases} q_{1-\alpha}(\gamma)/R(\gamma) & \textit{if } R > 0 \\ 0 & \textit{if } R = 0 \end{cases} .$$

The dependence on the choice of the rejection region $[0, \gamma]$ is omitted where possible for sake of notational simplicity.

Discoveries-at-Risk[1] is a random variable. Its realized value, for a given dataset, is then of the form in (2).

Proposition 1 *The realized value of DaR^{α} is, for a positive number of rejections, equal to*

$$\min_{\beta}\{\beta : P[V > \beta R_{obs}] \leq \alpha\}.$$

A proof is given in the Appendix.
The proportion of false discoveries is larger than the Discoveries-at-Risk value only with probability α. In other words, the interval $[0, DaR^{\alpha}]$ is a (1-α)-confidence interval for the proportion of false rejections.

Proposition 2 *In case R is strictly positive,*

$$P[V/R > DaR^{\alpha}] \leq \alpha.$$

In case R is not strictly positive, let Q be defined as in (1). Then $P[Q > DaR^{\alpha}] \leq \alpha$.

A proof is given in the Appendix.
The risk of underestimating the true proportion of false discoveries can thus be limited with Discoveries-at-Risk to any desired value. This is in sharp contrast to the false discovery rate, where such underestimation occurs frequently and cannot be controlled.

2.2 Controlling Discoveries-at-Risk.

Instead of working with a fixed rejection region and estimating the Discoveries-at-Risk value it might be of interest to reject as many hypotheses as possible while keeping the percentage of discoveries that are "at risk" below a certain value. In other words, one might want to find the largest possible rejection region such that the Discoveries-at-Risk value DaR^{α} does not exceed a certain threshold.

Definition 2 (Rejection Region) *The rejection region $\Gamma^{\alpha}_{DaR}(\beta) = [0, \gamma^{\alpha}_{DaR}(\beta)]$ is the largest rejection region such that the Discoveries-at-Risk value is controlled at value β:*

$$\gamma^{\alpha}_{DaR}(\beta) = \max_{\gamma \in [0,1]}\{\gamma : DaR^{\alpha}(\gamma) \leq \beta\}.$$

One can limit the percentage of discoveries at risk to any desired value through the choice of β. See e.g. Benjamini and Hochberg (1995), Storey et al. (2002) or Westfall and Young (1993) for similar approaches to FDR- and $FWER$-control respectively.

2.3 The relation to $FWER$

The Discoveries-at-Risk approach can be viewed as a generalization of the family-wise error rate ($FWER$).
Controlling the family-wise error rate at level α ensures that[2]

$$P[V/R > 0] \leq \alpha. \tag{3}$$

[1] Due to similar underlying ideas, the term "Discoveries-at-Risk" is intentionally chosen similar to Value-at-Risk, a well established method for financial risk management, see e.g. Duffie and Pan (1997).
[2] R is assumed to be strictly positive. If R is not strictly positive, (3) and (4) still hold if V/R is replaced by Q, as defined in (1).

Controlling DaR^α at value β ensures on the other hand that

$$P[V/R > \beta] \leq \alpha. \tag{4}$$

Controlling DaR^α at the value $\beta = 0$ is equivalent to controlling $FWER$ at level α. Equally many or more rejections occur if a small fraction of false rejections is accepted and DaR^α is controlled at some value $\beta > 0$.

Proposition 3 *Let $\Gamma_{FWER} = [0, \gamma^\alpha_{FWER}]$ be the rejection region such that FWER is controlled at level α (analogously to Definition 2). The rejection region $\Gamma_{DaR} = [0, \gamma^\alpha_{DaR}(\beta)]$ is for any value of β at least as large as the rejection region $\Gamma_{FWER} = [0, \gamma^\alpha_{FWER}]$.*

$$\gamma^\alpha_{FWER} = \gamma^\alpha_{DaR}(0) \leq \gamma^\alpha_{DaR}(\beta) \qquad \forall\, 0 \leq \beta \leq 1$$

A proof is given in the Appendix.
In an exploratory setting this property is quite useful as sometimes control of $FWER$ produces already sufficiently many rejections and one is certain to capture all of them when controlling DaR^α at any value β (which is not the case for *FDR*-control, see Theorem 3).
The main problem with $FWER$ is vanishing power for large numbers of tested hypotheses.
This holds under a fairly large class of dependency structures of the test statistics, as specified below.

Assumption 1 *The dependency between the test statistics is such that*

$$\begin{aligned} Var[V]/m_0^2 &\to 0 \qquad \textit{for}\, m_0 \to \infty, \\ Var[S]/m_1^2 &\to 0 \qquad \textit{for}\, m_1 \to \infty. \end{aligned}$$

Proposition 4 *Assumption 1 is fulfilled e.g. for*

1. *independent test statistics*
2. *correlated test statistics under the condition that the correlation matrix of $1_{[p_i \in \Gamma]}$ has block structure and the size l_{max} of the largest block is allowed to grow but not as fast as m_0:*

   $$l_{max}/m_0 \to 0 \qquad \textit{for}\, m_0 \to \infty.$$

   "Block structure" is here equivalent to

   $$Cov[1_{[\mathbf{p}_i \in \Gamma]}, 1_{[\mathbf{p}_j \in \Gamma]}] = 0$$

   for i, j in different blocks or partition cells of $\{1, \ldots, m\}$

A proof is given in the Appendix.

Theorem 1 *Under Assumption 1,*

$$\gamma^\alpha_{FWER} \to 0 \qquad \textit{for}\, m_0 \to \infty.$$

A proof is given in the Appendix.
The power of a $FWER$ controlling procedure thus vanishes for large numbers of tested hypotheses.
On the other hand, controlling DaR^α at small but strictly positive values has the important advantage that the power does *not* vanish for many tested hypotheses.
We make three reasonable assumptions.

Assumption 2 *For some* $0 < \pi_0 < 1$,

$$m_0 = \lfloor \pi_0 m \rfloor.$$

Note that the following results are valid as well if the number of true null hypotheses is e.g. binomially distributed as in Storey et al. (2002).

Assumption 3 *There exist real-valued functions* $F_0(\gamma)$, $F_1(\gamma)$, *strictly positive for* $\gamma > 0$, *such that*

$$E[S]/m_1 \rightarrow F_1(\gamma) \qquad \text{for}\, m_1 \rightarrow \infty, \tag{5}$$
$$E[V]/m_0 \rightarrow F_0(\gamma) = \gamma \qquad \text{for}\, m_0 \rightarrow \infty. \tag{6}$$

The second part of Assumption 3, (6), is just included for notational convenience as it is always fulfilled. Note that

$$E[V]/m_0 = \gamma = F_0(\gamma)$$

is indeed the cumulative distribution function of the uniformly distributed p-values under the null hypothesis and does not depend on the choice of m_0.
Assume that the specific form of alternative hypotheses is completely characterized by some vector ν. The first part of Assumption 3, (5), is then for example fulfilled if ν is independently drawn for each alternative hypothesis from some underlying distribution.
Assumptions 1 and 3 together only imply convergence in probability of S/m_1 and V/m_0 to F_1 and F_0 respectively. The assumptions are thus weaker than the assumption of almost sure pointwise convergence in Storey et al. (2002).

Assumption 4 F_0 *and* F_1 *are as specified in Assumption 3 and*

$$\frac{F_0(\gamma)}{F_1(\gamma)} \rightarrow 0 \qquad \text{for}\, \gamma \rightarrow 0.$$

For Assumption 4 to be valid, the function F_1 has to vanish more slowly for $\gamma \rightarrow 0$ than $F_0(\gamma) = \gamma$, the cumulative distribution function of uniformly distributed p-values under the alternative hypothesis. Note that, under Assumption 2,

$$\frac{F_0(\gamma)}{F_1(\gamma)} = c \cdot \lim_{m \rightarrow \infty} \frac{E[V]}{E[S]},$$

with $c = \pi_1/\pi_0$. Assumption 4 thus essentially states that we expect to find an overwhelming majority of true alternative hypotheses at infinitesimal small p-values.

Theorem 2 *Under Assumptions 1-4, there exists for every positive value of* β *a rejection region* $\Gamma = [0, \gamma]$ *with* $\gamma > 0$ *such that for large values of* m, $[0, \gamma^{\alpha}_{DaR}(\beta)]$ *contains* Γ *with arbitrarily high probability. That is,*

$$\forall \beta > 0\; \exists \gamma > 0: \qquad P[\gamma^{\alpha}_{DaR}(\beta) \geq \gamma] \rightarrow 1 \qquad \text{for}\, m \rightarrow \infty.$$

A proof is given in the Appendix.
Note that Theorem 1 is still valid under the same assumptions.

The advantage of allowing even a tiny percentage of false discoveries (by controlling DaR^α at strictly positive values β) is thus that the power of the resulting procedure does not vanish for large numbers of tested hypotheses.
Discoveries-at-Risk is hence a more flexible approach to multiple testing than *FWER*. The family-wise error rate can still be controlled in the Discoveries-at-Risk framework. But, if required, more powerful testing is possible by allowing a certain proportion $\beta > 0$ of false rejections.

2.4 The relation to *FDR*

There is a great risk that *FDR* underestimates the actual amount of false discoveries. The fact that Discoveries-at-Risk is capable of limiting this risk to any desired and typically low value α was shown in Proposition 2.
One might object, though, that there is "no free lunch" and a price has to be paid in order to achieve this better protection. This "price" might e.g. consist of fewer rejections if both DaR^α and *FDR* are controlled at the same value β. This is often true, though the difference was never large in the data we have seen, see also section 4.
Surprisingly, also the opposite behaviour regarding rejection power can occur. No rejections can be made if *FDR* is controlled at sufficiently low values.

Theorem 3 *Let $\Gamma_{FDR} = [0, \gamma_{FDR}(\beta)]$ be the rejection region such that FDR is controlled at value β. Not a single p-value $\mathbf{P}_i$, $i = 1, \ldots, m$ is in the rejection region $\Gamma_{FDR} = [0, \gamma_{FDR}(\beta)]$ if β is smaller than a strictly positive random variable β_0. That is*

$$\exists \beta_0 : \min_{1 \le i \le m} \{\mathbf{P}_i\} \notin [0, \gamma_{FDR}(\beta)] \qquad \forall\, \beta \le \beta_0.$$

A proof is given in the Appendix.
Assume that control of *FWER* at level α produces $r > 0$ rejections in a given dataset. If r is small, control of *FDR* might be interesting as false rejections are allowed and, hence, more rejections could be expected for control of *FDR* at some positive error rate than for control of *FWER*. But, maybe surprisingly, control of *FDR* will lead to zero rejections if a very low error rate is chosen; see Theorem 3. This cannot happen with control of DaR^α. For any error rate (including zero), there will be at least r rejections, as argued in Proposition 3. Thus DaR^α is more powerful than *FDR* when controlling at very low error rates. See section 4 for examples.
For very large numbers of tested hypotheses, however, it can be shown that both DaR^α and *FDR* give similar results.

Theorem 4 *For any fixed rejection region, DaR^α converges in probability to FDR under Assumptions 1-3,*

$$DaR^\alpha \xrightarrow{p} FDR \qquad \text{for } m \to \infty.$$

A proof is given in the Appendix.
It has to be noted, though, that even for several thousands of tested hypotheses the difference between DaR^α and *FDR* can still be large, particularly in case of dependent tests.
For larger rejection regions and dependent test statistics, the distribution of V/R is usually spread out (see section 4 for numerical examples) and the proportion of false rejections is often substantially higher than indicated by *FDR*. DaR^α is typically larger than *FDR* in such cases as it protects against frequent underestimation of V/R.

2.5 The role of the level α and value β

It might be confusing at first that two "error rates", namely α and β, have to be specified for control of DaR^{α}.
The error rate β measures the proportion of false rejections. The value of β thus serves as the maximal value of DaR^{α} or *FDR* that one is willing to accept. Control of *FWER* implicitly corresponds to the choice $\beta = 0$ in DaR^{α}, as shown in section 2.3.
The true proportion of false discoveries is, however, not bounded by β. The probability that the true proportion of false discoveries is larger than β is measured by α. This probability can be controlled at suitable low values in either the DaR^{α}- or *FWER*-approach to multiple testing. As mentioned above and illustrated in section 4, this second error rate α cannot be controlled and is usually large in the *FDR*-approach to multiple testing.
Our explicit notation should help to distinguish between these two different error rates.

3 Estimating Discoveries-at-Risk

First we propose an estimate of Discoveries-at-Risk for a fixed rejection region. Second, it is shown how this estimate can be used for control of Discoveries-at-Risk.

3.1 Estimating Discoveries-at-Risk for a fixed rejection region

By definition, the value of DaR^{α} is given by the distribution of V. We estimate the distribution of V under the complete null hypothesis, assuming that all hypotheses are true null hypotheses. A similar approach was taken in Benjamini and Hochberg (1995) for control of the false discovery rate or in Westfall and Young (1993) for single-step control of *FWER*.

Definition 3 *Let $\mathbf{P}^c$ be the random vector of p-values under the complete null hypothesis. The random variable V^c is defined as the number of rejected hypotheses under the complete null hypothesis,*

$$V^c(\gamma) = \sum_{i=1}^{m} 1_{[\mathbf{P}_i^c \in \Gamma]}.$$

Definition 4 *Let $q^c_{1-\alpha}(\gamma)$ be the (1-α)-quantile of the distribution of $V^c(\gamma)$. DaR^{α} is estimated as*

$$\widehat{DaR}^{\alpha}(\gamma) = \begin{cases} q^c_{1-\alpha}(\gamma)/R(\gamma) & \text{if } R(\gamma) > 0 \\ 0 & \text{if } R(\gamma) = 0 \end{cases}$$

The proposed estimate provides strong control over DaR^{α} in that it is larger than the true value of DaR^{α}.

Theorem 5 *It holds for all possible realizations that*

$$\widehat{DaR}^{\alpha}(\gamma) \geq DaR^{\alpha}(\gamma).$$

Furthermore, for all $\gamma \geq 0$, $P[V/R > \widehat{DaR}^{\alpha}] \leq \alpha$.

A proof is given in the Appendix.

3.2 Estimating Rejection Regions

The estimate of the rejection region follows immediately by replacing in Definition 2 the value of DaR^α by the proposed estimate $\widehat{DaR}^\alpha$.

Definition 5 *We estimate the rejection region* $\Gamma = [0, \gamma^\alpha_{DaR}(\beta)]$ *by:*

$$\widehat{\gamma}^\alpha_{DaR}(\beta) = \max_{\gamma \in [0,1]} \{\gamma : \widehat{DaR}^\alpha(\gamma) \leq \beta\}.$$

Corollary 1 *The estimated rejection region is conservative. For all possible realizations it holds that*

$$\widehat{\gamma}^\alpha_{DaR}(\beta) \leq \gamma^\alpha_{DaR}(\beta) \qquad \forall\, 0 < \alpha < 1\,, \beta \geq 0.$$

This is a direct consequence of Definition 5 and Theorem 5.
As stated above, controlling DaR^α at value $\beta = 0$ is equivalent to controlling $FWER$ at level α. It is thus of interest to compare the power of controlling DaR^α at value $\beta = 0$ (as proposed in Definition 5) to other $FWER$-controlling procedures.

Theorem 6 *Controlling the Discoveries-at-Risk value with the proposed estimate* $\widehat{DaR}^\alpha$ *at* $\beta = 0$ *achieves the same power (the same number of rejections) for controlling FWER as the single-step minP controlling procedure of Westfall and Young (1993).*

A proof is given in the Appendix.
Controlling DaR^α at value $\beta = 0$ is thus a slightly less powerful $FWER$-controlling procedure than the step-down method of Westfall and Young (1993). We emphasize that this is just a property of the proposed *estimate* of DaR^α. Finding less conservative estimates of DaR^α is clearly desirable. We note, though, that almost no difference in power is visible between the two mentioned $FWER$-controlling procedures in the application of detecting differential gene expression from microarray data, see the following section.

4 Numerical Studies

We demonstrate the usefulness of the Discoveries-at-Risk framework both with simulated and real data and compare results to existing methods.

4.1 Simulated Data

A two-class problem with 50 different observations $\mathbf{Z}^k \in \mathbb{R}^m$, $k = 1, \ldots, 50$, is constructed with $m = 5000$:

$$\begin{array}{lll} \text{class 1:} & \mathbf{Z}^1, \ldots, \mathbf{Z}^{25} & i.i.d. \sim \mathcal{N}(0, \Sigma) \\ \text{class 2:} & \mathbf{Z}^{26}, \ldots, \mathbf{Z}^{50} & i.i.d. \sim \mathcal{N}(\mu, \Sigma), \\ \text{with} & \mu_i = \left\{ \begin{array}{ll} 0 & i \in \mathcal{A}_0 \\ 1 & i \in \mathcal{A}_1 \end{array} \right. , & \end{array}$$

where $\mathcal{A}_0$ is a random subset of $\{1, \ldots, m\}$ with cardinality $|\mathcal{A}_0| = 4500$ and $\mathcal{A}_1$ is its complement. Using the two-sample Wilcoxon test at individual significance level 5%, we

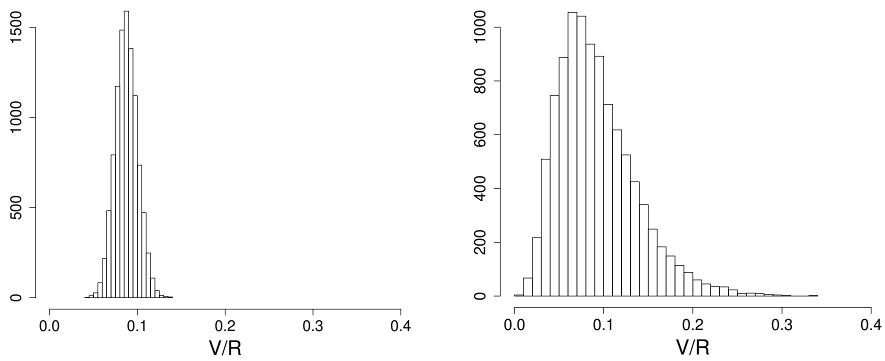

Figure 1: The distribution of V/R for $l = 1$, corresponding to independent test statistics (left), and block dependent test statistics with block-size $l = 500$ (right).

test for each component i, $i = 1, \ldots, m$ whether the distribution of $\mathbf{Z}_i^1, \ldots, \mathbf{Z}_i^{25}$ is different from the distribution of $\mathbf{Z}_i^{26}, \ldots, \mathbf{Z}_i^{50}$.
The distribution is, in reality, the same for 4500 hypotheses as $\mu_i = 0$ if the i-th null hypothesis is true. The distribution is shifted between the two classes by setting $\mu_i = 1$ for the remaining 500 components.
We assume that the dependency between the components from different hypotheses has block-structure, see Proposition 4. The correlation between hypotheses within one block is moderate at 0.25. All blocks have equal size l, chosen for notational simplicity such that m is a multiple of l. We denote the set of hypotheses in one block by $\mathcal{L}_k$, $k = 1, \ldots, m/l$. Then

$$\Sigma_{i,j} = \begin{cases} 1 & i = j \\ 0.25 & i \neq j,\ \exists k : i, j \in \mathcal{L}_k \\ 0 & \text{otherwise} \end{cases} .$$

In Figure 1, the histograms of the proportion V/R of false rejections[3] are shown for 10'000 simulations. The variance of V/R is clearly larger for positively correlated test statistics than for independent test statistics.
In Table 2 the probability of underestimating V/R with either FDR or $DaR^{0.05}$ is shown. While V/R is larger than $DaR^{0.05}$ in less than 5% of all simulations (as argued in Proposition 2), V/R is larger than FDR in between one-third and one-half of all cases. The expected shortfall is largest for dependent test statistics: if V/R is e.g. larger than FDR in the case of $l = 5000$, it is larger on average by 138%,

$$E\Big[\frac{V/R}{FDR}\Big|V/R > FDR\Big] = 2.38.$$

[3]There has never been a problem with the quotient V/R, as in all simulations the number of rejections has been strictly positive. For a more formal, but for practical purposes equivalent, definition one might use the definition Q as in (1).

	$l=1$	$l=500$	$l=5000$
$P[V/R > FDR]$	49.7%	43.2%	32.2%
$P[V/R > DaR^{\alpha}]$	4.6%	4.7%	4.8%
$E[FDR/DaR^{\alpha}]$	0.81	0.46	0.20

Table 2: The risk (in percent) of underestimating V/R with FDR and $DaR^{0.05}$.

	$l=1$	$l=500$	$l=5000$
$P[V/R > \widehat{FDR}^{bh}]$	27.9%	35.1%	23.3%
$P[V/R > \widehat{FDR}^{st}]$	55.6%	46.5%	31.5%
$P[V/R > \widehat{DaR}^{0.05}]$	0%	1.9%	4.0%
$E[\widehat{FDR}^{bh}/\widehat{DaR}^{0.05}]$	0.69	0.45	0.26
$E[\widehat{FDR}^{st}/\widehat{DaR}^{0.05}]$	0.62	0.41	0.24

Table 3: The risk (in percent) of underestimating V/R with the Benjamini-Hochberg estimate $\widehat{FDR}^{bh}$ of FDR, the Storey estimate $\widehat{FDR}^{st}$ and the proposed estimate $\widehat{DaR}^{0.05}$ of the Discoveries-at-Risk value $DaR^{0.05}$.

In Table 3 it can be seen that the proportion V/R of false rejections is frequently underestimated not only by the true value of FDR but as well by *estimates* of FDR. This probability is shown for the Benjamini-Hochberg estimate of FDR,

$$\widehat{FDR}^{bh} = \frac{m\gamma}{R_{obs}} \tag{7}$$

and the Storey-estimate $\widehat{FDR}^{st}$ of FDR, setting $\lambda = 0.5$. For details see Storey (2002). Although the latter estimate is on average larger than the true value of FDR, it achieves even less protection against underestimation of V/R than the true value of FDR (at least for independent test statistics). This is due to the fact that this estimate is negatively correlated with V/R.
The proposed estimate of DaR^{α} on the other hand achieves the argued protection against underestimation of V/R (see Theorem 5).
It has thus been demonstrated that the risk of underestimating V/R with FDR is very large. This risk can easily be controlled at any desirable value with the Discoveries-at-Risk approach.

4.2 Detection of differential gene expression in microarray data

With microarray experiments it is possible to monitor expression levels in cells for several thousands of genes simultaneously (Alon et al. (1999); Golub et al. (1999); Ross et al. (2000)).
A common aim is to identify genes that are differentially expressed, that is associated with a variable of interest, such as tumour subtypes. We looked at three publicly available datasets, where the variable is a binary class label, distinguishing either between normal

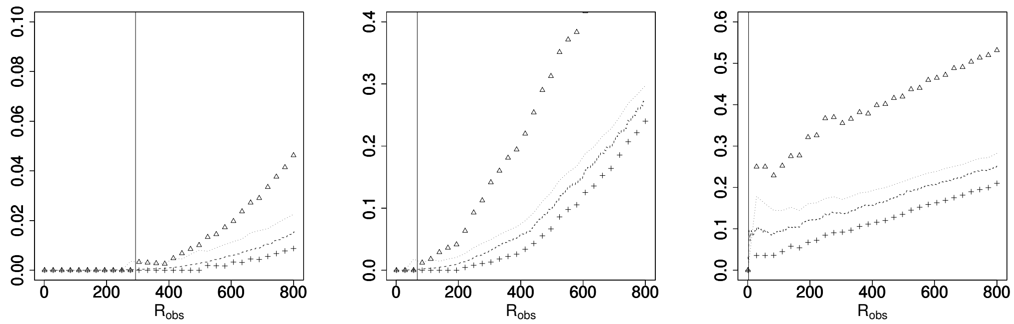

Figure 2: Various error rates of false discoveries as a function of the number R_{obs} of rejected hypotheses: (a, '–') the Benjamini-Hochberg estimate of FDR, (b,'+') the estimate of $DaR^{0.5}$, (c,'$\triangle$') the estimate of $DaR^{0.05}$ and (d,'...') the estimate of $DaR^{0.05}$ under assumed independence between the test statistics. The number of hypotheses that can be rejected when controlling $FWER$ at level $\alpha = 0.05$ is indicated by a vertical line and corresponds to the region where the estimate of $DaR^{0.05}$ is identically zero. The data are taken from leukemia (left), colon (middle) and breast cancer studies (right).

and tumourous colon tissue (Alon et al. (1999)), two subtypes of leukemia (Golub et al. (1999)) or clinical outcome for breast cancer (van't Veer et al. (2002)).

Preprocessing of the data is in each case performed as suggested by the authors. For each gene we test if the expression levels are stochastically larger for one or the other class using the Wilcoxon-test.

Under the complete null hypothesis, the distribution of the number of falsely rejected hypotheses is obtained by permutations of the class labels, see Dudoit et al. (2000) and Ge et al. (2003). This yields the distribution of V^c (see section 3.1), conditional on the observed expression matrix and conditional on the number of members in each class. The number of randomly sampled permutations is restricted here to 20'000 for each dataset.

Comparison with $FWER$. Controlling the estimate of DaR^{α} at value $\beta = 0$ was shown in Theorem 6 to be equivalent to the single-step resampling based $FWER$-controlling procedure of Westfall and Young (1993). The resulting number of rejections are shown in the first column of Table 4 and are indicated by a vertical line in Figure 2.

We compare the power to other $FWER$-controlling procedures. The number of possible rejections of the arguably most powerful method for strong control of $FWER$, the step-down procedure by Westfall and Young (1993), is shown in the second column of Table 4. Furthermore we give in the third column the number of rejections under the step-down $FWER$-controlling procedure of Holm (1979) and, in the fourth column, the number of rejections when Bonferronis correction is applied. A comprehensive review of these methods in the context of testing differential gene expression in microarray data can be found in Ge et al. (2003).

The proposed DaR^{α}-controlling procedure achieves on all three datasets almost the same power as the resampling-based step-down $FWER$-controlling procedure of Westfall and Young. Less powerful are Holm's step-down method and Bonferronis correction, the best

		$R\big(\widehat{\gamma}^{\alpha}_{DaR}(0)\big)$	W. and Y. step-down	Holm step-down	Bonferroni correction
leukemia	$\alpha = 0.01$	196	196	193	191
	$\alpha = 0.05$	281	288	269	266
	$\alpha = 0.1$	332	339	312	307
colon	$\alpha = 0.01$	36	36	32	32
	$\alpha = 0.05$	68	68	55	55
	$\alpha = 0.1$	90	92	69	69
breast	$\alpha = 0.01$	1	1	0	0
	$\alpha = 0.05$	3	3	2	2
	$\alpha = 0.1$	3	3	3	3

Table 4: Number of rejected hypotheses for various $FWER$-controlling procedures and levels α.

		$R\big(\widehat{\gamma}^{0.05}_{DaR}(\beta)\big)$	$R\big(\widehat{\gamma}^{0.05}_{DaR}(\beta)\big)$	$R\big(\widehat{\gamma}^{st}_{FDR}(\beta)\big)$	$R\big(\widehat{\gamma}^{bh}_{FDR}(\beta)\big)$
leukemia	$\beta = 0$	281	501	0	0
	$\beta = 0.01$	509	831	701	855
	$\beta = 0.05$	811	1182	1093	1360
	$\beta = 0.1$	1025	1407	1346	1694
colon	$\beta = 0$	68	203	0	0
	$\beta = 0.01$	100	274	194	241
	$\beta = 0.05$	194	460	373	472
	$\beta = 0.1$	253	563	492	641
breast	$\beta = 0$	3	20	0	0
	$\beta = 0.01$	3	20	0	0
	$\beta = 0.05$	3	123	3	3
	$\beta = 0.1$	3	344	28	323

Table 5: Number of rejections if Discoveries-at-Risk and the false discovery rate are controlled at various values β.

known but least powerful of all multiple testing procedures. The latter two methods do not take the dependency between the test statistics properly into account.

Comparison with *FDR*. The number of rejected hypotheses for DaR^{α}-control and FDR-control at various values of β is shown in Table 5. More specifically, the number of rejections is shown for control of the estimate of $DaR^{0.05}$ in the first and $DaR^{0.5}$ in the second column. The number of rejections for FDR-control with the Benjamini-Hochberg estimate (7) of FDR is shown in the third and with the Storey-estimate of FDR in the fourth column (setting $\lambda = 0.5$, see Storey (2002)).
It can be observed in Table 5 that the number of rejected hypotheses vanishes if FDR is controlled at small values β (as argued already in Theorem 3). Comparing with Table 4, it can furthermore be seen that controlling DaR^{α} at any value $\beta \geq 0$ leads to at least as many rejections as controlling $FWER$ at level α (see as well Proposition 3).
Controlling at the positive but small value $\beta = 0.01$, more rejections occurred for control

of $DaR^{0.05}$ than for control of FDR in the most difficult data (breast cancer), but less in the easier cases of leukemia and colon cancer data. For larger values of β, control of FDR leads typically to more rejections than control of $DaR^{0.05}$. Put differently, the estimate of $DaR^{0.05}$ is greater than the estimate of FDR for larger rejection regions. This is mostly due, however, to the high degree of dependency between the test statistics and the resulting high variance of the number of false rejections. For comparison, the estimate of $DaR^{0.05}$ under assumed independence between the test statistics is shown in Figure 2. Hence, as already argued by Figure 1 and Tables 2 and 3, the power of FDR is likely to be caused by not properly protecting against type I errors.

5 Conclusions

Controlling the Discoveries-at-Risk value DaR^{α} is a new and flexible tool in multiple hypothesis testing. It offers substantial advantages compared to the common control of the family-wise error rate ($FWER$) or the false discovery rate (FDR).

If desired, $FWER$ can still be controlled in the Discoveries-at-Risk framework but more powerful procedures are possible by allowing a certain fraction of false discoveries. This is in common with the false discovery rate. But the true amount of false discoveries is often very much larger than indicated by FDR. This shortcoming of FDR is corrected in the Discoveries-at-Risk approach. Surprisingly, besides the tight control of underestimating the amount of false discoveries, the DaR^{α} approach has even more power than FDR when controlling at very low error rates.

The new Discoveries-at-Risk approach to multiple hypotheses testing thus combines the advantages of both the false discovery and family-wise error rate while avoiding poor power (the major problem with $FWER$) on the one hand and low protection against underestimation of the true amount of false discoveries (as with FDR) on the other hand.

References

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology 96*, 6745–6750.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica 12*(1), 111–139.

Duffie, D. and J. Pan (1997). An overview of value at risk. *Journal of Derivatives*, 7–49.

Ge, Y., S. Dudoit, and T. Speed (2003). Resampling-based multiple testing for microarray data analysis. Technical report, Department of Statistics, University of California, Berkeley.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caliguri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Ross, D. T., U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. van der Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein, and P. O. Brown (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics 24*, 227–234.

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of The American Statistical Association 84*(406), 608–610.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B 64*(3), 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2002). A unified estimation approach to false discovery rates. Technical report, Department of Statistics, University of California, Berkeley.

van't Veer, L. J., H. Dal, M. J. van der Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 406*, 742–747.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons.

6 Appendix: Proofs

Proof of Proposition 1. For a positive number of rejections R_{obs}, the realization of DaR^α is equal to

$$\begin{aligned} DaR^\alpha &= q_{1-\alpha}/R_{obs} \\ &= \min_n \{n : P[V > n] \leq n\}/R_{obs} \\ &= \min_\beta \{\beta : P[V > \beta R_{obs}] \leq \alpha\}. \end{aligned}$$

Proof of Proposition 2. It suffices to show the second claim. By definition of Q, equation (1),

$$\begin{aligned} P\big[Q(\gamma) > DaR^\alpha(\gamma)\big] &= P[\frac{V(\gamma)}{R(\gamma)} > \frac{q_{1-\alpha}(\gamma)}{R(\gamma)} \big| R(\gamma) > 0]\ P\big[R(\gamma) > 0\big] \\ &\leq P[V(\gamma) > q_{1-\alpha}(\gamma)] \\ &\leq \alpha, \end{aligned}$$

as $q_{1-\alpha}(\gamma)$ is the $1-\alpha$ quantile of $V(\gamma)$.

Proof of Proposition 3. Let $\Gamma = [0, \gamma^{\alpha}_{FWER}]$ be the rejection region for controlling $FWER$ at level α. Then

$$\begin{array}{rrcll} & P[V(\gamma)>0] & \leq & \alpha & \forall\gamma \leq \gamma^{\alpha}_{FWER}, \\ \text{and} & P[V(\gamma)>0] & > & \alpha & \forall\gamma \geq \gamma^{\alpha}_{FWER}. \end{array}$$

Hence $q_{1-\alpha}(\gamma) = 0$ for all $\gamma \in \Gamma$ and $q_{1-\alpha}(\gamma) > 0$ for all $\gamma \notin \Gamma$. By Definition 1,

$$DaR^{\alpha}(\gamma)\begin{cases} =0 & \gamma \leq \gamma^{\alpha}_{FWER}, \\ >0 & \gamma \geq \gamma^{\alpha}_{FWER}. \end{cases}$$

It then follows by Definition 2 that $\gamma^{\alpha}_{DaR}(0) = \gamma^{\alpha}_{FWER}$.
Furthermore, by Definition 2,

$$\gamma^{\alpha}_{DaR}(\beta) \geq \gamma^{\alpha}_{DaR}(0) \qquad \forall\beta > 0.$$

This completes the proof of Proposition 3.

Proof of Proposition 4. It is sufficient to show claim (b) for V. We assume for notational simplicity that m_0 is a multiple of l_{max},

$$\exists k \in \mathbb{N}: \qquad kl_{max} = m_0.$$

The variance of V is bounded by

$$Var[V] \leq l^2_{max} \cdot k = m_0 \cdot l_{max}.$$

Hence

$$Var[V]/m_0^2 \leq l_{max}/m_0 \to 0 \qquad \text{for } m_0 \to \infty.$$

Proof of Theorem 1. For any rejection region $\Gamma = [0, \gamma]$, $FWER$ is only controlled at a given level α if it holds that

$$P[V(\gamma) = 0] \geq 1 - \alpha.$$

For a proof of the proposition it is hence sufficient to show that for any rejection region $\Gamma = [0, \gamma]$,

$$P[V(\gamma) = 0] \to 0 \qquad \text{for } m_0 \to \infty. \tag{8}$$

To show (8), some $0 < \kappa < 1$ is chosen. Then

$$\begin{array}{rcl} P[V(\gamma)=0] & \leq & P[|V(\gamma) - m_0\gamma| = m_0\gamma] \\ & \leq & P[|V(\gamma) - m_0\gamma| > \kappa m_0\gamma]. \end{array}$$

Using the Chebyshev inequality,

$$P[|V(\gamma) - m_0\gamma| > \kappa m_0\gamma] \leq \frac{Var[V(\gamma)]}{\kappa^2 m_0^2 \gamma^2}.$$

Under Assumption 1 and any fixed values of γ and κ, the last quantity vanishes for $m_0 \to \infty$. This proves (8).

Lemma 1 *Under Assumptions 1-3 it holds pointwise that*

$$V(\gamma)/m \quad \xrightarrow{p} \quad \pi_0 F_0(\gamma) \qquad for\, m \to \infty, \tag{9}$$
$$R(\gamma)/m \quad \xrightarrow{p} \quad \pi_0 F_0(\gamma) + \pi_1 F_1(\gamma) \qquad for\, m \to \infty, \tag{10}$$

where F_0 and F_1 are defined as in Assumption 3.

Proof of Lemma 1. To prove the first claim (9), it is sufficient to show that under Assumptions 2 and 3,

$$\frac{E[V(\gamma)]}{m} \to \pi_0 F_0(\gamma) \qquad \text{for } m \to \infty. \tag{11}$$

Formula (9) follows then by the Chebyshev inequality and Assumption 1. But

$$\frac{E[V(\gamma)]}{m} = \frac{m_0}{m}\frac{E[V(\gamma)]}{m_0}.$$

By Assumption 2,

$$\frac{m_0}{m} \to \pi_0 \qquad \text{for } m \to \infty.$$

Furthermore,

$$\frac{E[V(\gamma)]}{m_0} = \gamma = F_0(\gamma).$$

This implies (11) and hence (9).
It remains to show the second claim (10). First,

$$E[\frac{R(\gamma)}{m}] = E[\frac{S(\gamma)+V(\gamma)}{m}] = \frac{m_1}{m}\frac{E[S(\gamma)]}{m_1} + \frac{m_0}{m}\frac{E[V(\gamma)]}{m_0}.$$

It was shown above that

$$\frac{m_0}{m}\frac{E[V(\gamma)]}{m_0} \to \pi_0 F_0(\gamma) \qquad \text{for } m \to \infty..$$

Similarly, it follows by Assumptions 2 and 3 that

$$\frac{m_1}{m}\frac{E[S(\gamma)]}{m_1} \to \pi_1 F_1(\gamma) \qquad \text{for } m \to \infty,$$

which completes the Proof of Lemma 1.

Lemma 2 *Under Assumptions 1-4, there exists a function $g(\gamma)$ with the properties*

$$\gamma/g(\gamma) \to 0 \qquad for\, \gamma \to 0 \tag{12}$$

such that it holds pointwise for all $\gamma > 0$ that

$$P[R(\gamma) < mg(\gamma)] \to 0 \qquad for\, m \to \infty. \tag{13}$$

Proof of Lemma 2. We claim that the function

$$g^\star(\gamma) = \frac{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)}{2}$$

has the desired property. Properties (12) and (13) have to be verified.
Using $F_0(\gamma) = \gamma$ and observing that $\pi_1 > 0$ according to Assumption 2,

$$\gamma/g^\star(\gamma) = \frac{2F_0(\gamma)}{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)} \leq \frac{2}{\pi_1}\frac{F_0(\gamma)}{F_1(\gamma)}$$

By Assumption 4,

$$\frac{F_0(\gamma)}{F_1(\gamma)} \to 0 \qquad \text{for } \gamma \to 0.$$

Hence

$$\gamma/g^\star(\gamma) \to 0 \qquad \text{for } \gamma \to 0,$$

and property (12) is shown for $g^\star$.
It remains to show property (13). By Lemma 1, $R(\gamma)/m$ converges pointwise in probability to $2g^\star(\gamma)$. Hence

$$R(\gamma) \xrightarrow{p} 2mg^\star(\gamma) \qquad \text{for } m \to \infty.$$

Property (13) follows thus immediately for $g^\star(\gamma)$.

Proof of Theorem 2. Given $\beta > 0$, we show that there exists a $\gamma^\star > 0$ such that

$$P[DaR^\alpha(\gamma^\star) \leq \beta] \to 1 \qquad \text{for } m \to \infty. \tag{14}$$

If (14) holds, then

$$P[\gamma^\alpha_{DaR}(\beta) \geq \gamma^\star] \to 1 \qquad \text{for } m \to \infty,$$

and the claim is proven.
Using the Chebyshev inequality, it holds for any $c > 1/\sqrt{\alpha}$ that

$$\begin{aligned} P\big[V(\gamma) \notin [0, m_0\gamma + c\sqrt{Var[V(\gamma)]}\big] &\leq P\big[|V(\gamma) - m_0\gamma| > c\sqrt{Var[V(\gamma)]}\big] \\ &\leq \alpha. \end{aligned}$$

Hence

$$q_{1-\alpha}(\gamma) \leq m_0\gamma + c\sqrt{Var[V(\gamma)]}.$$

By Definition 1, it follows that

$$\begin{aligned} DaR^\alpha(\gamma) &\leq \begin{cases} (m_0\gamma + c\sqrt{Var[V(\gamma)]})/R(\gamma) & R(\gamma) > 0 \\ 0 & R(\gamma) = 0 \end{cases} \\ &\leq \begin{cases} (m\gamma + c\sqrt{Var[V(\gamma)]})/R(\gamma) & R(\gamma) > 0 \\ 0 & R(\gamma) = 0 \end{cases} \end{aligned}$$

Under Assumptions 1-4, Lemma 2 states that there exists a function $g(\gamma)$ with property (12) such that

$$P[R(\gamma) < mg(\gamma)] \to 0 \qquad \forall \gamma > 0.$$

This implies that

$$\forall\, \epsilon > 0 \;\; \exists m' \in \mathbb{N} : \qquad P\Big[\frac{1}{R(\gamma)} \leq \frac{1}{mg(\gamma)}\Big| R(\gamma) > 0\Big] \geq 1 - \epsilon \qquad \forall m \geq m'.$$

Using this in (15), it follows that

$$\forall\, \epsilon > 0 \;\; \exists m' \in \mathbb{N} : \qquad P\Big[DaR^{\alpha}(\gamma) \leq \frac{\gamma}{g(\gamma)} + \frac{c\sqrt{Var[V(\gamma)]}}{mg(\gamma)}\Big] \geq 1 - \epsilon \qquad \forall m \geq m'. \quad (15)$$

Note that the value of m' might depend on both γ and ϵ.
As by (12), $\gamma/g(\gamma) \to 0$ for $\gamma \to 0$, there exists a $\gamma^{\star} > 0$ such that

$$\gamma^{\star}/g(\gamma^{\star}) \leq \beta/2. \quad (16)$$

Given c and $\gamma^{\star}$, it is possible under Assumptions 1 and 2 to choose $m^{\star} > m'$ such that

$$\frac{c\sqrt{Var[V(\gamma^{\star})]}}{mg(\gamma^{\star})} \leq \beta/2 \qquad \forall\, m \geq m^{\star}.$$

Putting this and (16) back into (15),

$$\forall\, \beta > 0, \epsilon > 0 \;\; \exists \gamma^{\star} > 0, m^{\star} \in \mathbb{N} : \qquad P[DaR^{\alpha}(\gamma^{\star}) \leq \beta] \geq 1 - \epsilon \qquad \forall\, m \geq m^{\star}, \quad (17)$$

which proves (14) and completes thus the proof of Theorem 2.

Proof of Theorem 3. For any rejection region $[0, \gamma]$ with $\gamma > 0$,

$$P[V(\gamma) = 0] < 1.$$

Let Q be defined as in equation (1). Then

$$P[Q(\gamma) = 0] < 1,$$

and hence $E[Q(\gamma)] > 0$. The value of FDR is thus strictly positive for any non-vanishing rejection region.
Let

$$\beta_0 = FDR(\gamma^{\star})$$

be the strictly positive value of the false discovery rate for the rejection region

$$[0, \gamma^{\star}] = [0, \min_{1 \leq i \leq m} \{\mathbf{P}_i\}],$$

where $\min_{1 \leq i \leq m}\{\mathbf{P}_i\}$ is the minimum of all p-values. No rejections occur if FDR is controlled at any value $\beta < \beta_0$.

Proof of Theorem 4. We show the Proposition in two parts. First, it is shown that

$$DaR^{\alpha}(\gamma) \xrightarrow{p} \frac{\pi_0 F_0(\gamma)}{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)} \qquad \text{for } m \to \infty. \tag{18}$$

Second, we show that

$$FDR(\gamma) \to \frac{\pi_0 F_0(\gamma)}{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)} \qquad \text{for } m \to \infty. \tag{19}$$

The claim follows from (18) and (19).
It suffices to proof the claim for strictly positive $R(\gamma)$ as in case $R(\gamma) = 0$,

$$DaR^{\alpha}(\gamma) = 0 = FDR(\gamma). \tag{20}$$

For strictly positive $R(\gamma)$,

$$DaR^{\alpha}(\gamma) = \frac{q_{1-\alpha}(\gamma)}{m} \frac{1}{R(\gamma)/m},$$

where $q_{1-\alpha}(\gamma)$ is the (1-α)- quantile of $V(\gamma)$.
It follows directly from the first claim in Lemma 1 and Assumptions 1 and 2 that

$$q_{1-\alpha}(\gamma)/m \to \gamma\pi_0 \qquad \text{for } m \to \infty. \tag{21}$$

Furthermore, it follows by Lemma 1 and Assumption 2 that

$$\frac{1}{R(\gamma)/m} \xrightarrow{p} \frac{1}{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)} \qquad \text{for } m \to \infty.$$

Together with (21), (18) follows and it remains to show (19).
Now,

$$\frac{V(\gamma)}{R(\gamma)} = \frac{V(\gamma)}{m} \frac{1}{R(\gamma)/m}$$

Hence, by Lemma 1,

$$\frac{V(\gamma)}{R(\gamma)} \xrightarrow{p} \frac{\pi_0 F_0(\gamma)}{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)} \qquad \text{for } m \to \infty.$$

As furthermore

$$\left|\frac{V(\gamma)}{R(\gamma)}\right| \leq 1 \qquad \forall \gamma \geq 0,$$

it follows that

$$E\left[\frac{V(\gamma)}{R(\gamma)}\right] \to \frac{\pi_0 F_0(\gamma)}{\pi_0 F_0(\gamma) + \pi_1 F_1(\gamma)} \qquad \text{for } m \to \infty.$$

This proves (19) and completes the proof of Theorem 4.

Proof of Theorem 5. Let h_i be equal to H_0 if the i-th hypothesis is a null hypothesis. If $h_i = H_0$, $\mathbf{P}_i^c = \mathbf{P}_i$ by definition of $\mathbf{P}_i^c$. Hence

$$V(\gamma) = \sum_{\substack{i=1 \\ h_i = H_0}}^{m} 1_{[\mathbf{P}_i \in \Gamma]} \leq \sum_{i=1}^{m} 1_{[\mathbf{P}_i^c \in \Gamma]} = V^c(\gamma).$$

Therefore

$$P\big[V(\gamma) > k\big] \leq P\big[V^c(\gamma) > k\big] \qquad \forall k \in \mathbb{N}, \gamma \geq 0.$$

and

$$q_{1-\alpha}^c(\gamma) \geq q_{1-\alpha}(\gamma) \qquad \forall \gamma > 0.$$

Hence, for all possible realizations,

$$\widehat{DaR}^{\alpha}(\gamma) \geq DaR^{\alpha}(\gamma) \qquad \forall \gamma > 0. \tag{22}$$

The second claim follows directly from (22) and Proposition 2.

Proof of Theorem 6. Let $\mathbf{p}_{(i)}$ be the ordered realizations of unadjusted p-values $\mathbf{P}_i$ in a multiple testing situation with

$$\mathbf{p}_{(1)} \leq \mathbf{p}_{(2)} \leq \cdots \leq \mathbf{p}_{(m)}.$$

Let $\tilde{\mathbf{p}}_{(i)}, i = 1, \ldots, m$ be the corresponding single-step, minP adjusted p-values as in Westfall and Young (1993),

$$\tilde{\mathbf{p}}_{(i)} = P[\min_{l=1,\ldots,m} \mathbf{P}_l^c \leq \mathbf{p}_{(i)}], \tag{23}$$

where $\mathbf{P}^c$ is the random variable of p-values under the complete null hypothesis.
For any rejection region $\Gamma = [0, \gamma]$, we have by definition of V^c that

$$P[V^c(\gamma) > 0] = P[\min_{l=1,\ldots,m} \mathbf{P}_l^c \leq \gamma]. \tag{24}$$

In case that no rejections occur for control of $FWER$ with the single-step method at level α, we have

$$P[V^c(\mathbf{p}_{(i)}) > 0] > \alpha \qquad \forall i = 1, \ldots, m$$

and hence $q_{1-\alpha}^c(\mathbf{p}_{(i)}) > 0$ for all $i = 1, \ldots, m$. Therefore no rejections occur as well for control of $\widehat{DaR}^{\alpha}$ at value $\beta = 0$.
For a positive number of rejections with the single-step method, the rejection region $[0, \widehat{\gamma}_{DaR}^{\alpha}(0)]$ for control of $\widehat{DaR}^{\alpha}$ at value $\beta = 0$ is given by

$$\begin{aligned} \widehat{\gamma}_{DaR}^{\alpha}(\beta = 0) &= \max\{\gamma : \widehat{DaR}^{\alpha}(\gamma) = 0\} \\ &= \max\{\gamma : q_{1-\alpha}^c(\gamma) = 0\} \\ &= \max\{\gamma : P[V^c(\gamma) > 0] \leq \alpha\} \\ &= \max\{\gamma : P[\min_{l=1,\ldots,m} \mathbf{P}_l^c \leq \gamma] \leq \alpha\}. \end{aligned} \tag{25}$$

It is clear by (23) that the rejection region defined by (25) includes those and only those test statistics with a single-step $FWER$-adjusted p-value $\tilde{\mathbf{p}}_{(i)}$ below or at α. This completes the proof of Theorem 6.