Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 4, *pp.* 405–416

Discovering the false discovery rate

Yoav Benjamini

Tel Aviv University, Israel

[*Presented to* The Royal Statistical Society *at its 175th-anniversary conference in a session organized by the* Research Section *on Wednesday, September 9th, 2009*, Professor D. Firth *in the Chair*]

Summary. I describe the background for the paper ‘Controlling the false discovery rate: a new and powerful approach to multiple comparisons’ by Benjamini and Hochberg that was published in the *Journal of the Royal Statistical Society*, Series B, in 1995. I review the progress since made on the false discovery rate, as well as the major conceptual developments that followed.

Keywords: False coverage rate; Multiple comparisons; ‘Testimation’

1. Background

Our work on the false discovery rate (FDR), and the paper Benjamini and Hochberg (1995), has its origins in two papers concerned with multiple testing of m hypotheses of which unknown m_0 are true. First was Schweder and Spjøtvoll (1982), who suggested plotting the ranked p-values, assessing $\mathbf{m}_0$ via an eye-fitted line, and rejecting the other $m-\mathbf{m}_0$ hypotheses. In Hochberg and Benjamini (1990) we developed their idea into an algorithm and incorporated the estimate $\mathbf{m}_0$ into procedures such as Bonferroni, Holm or Hochberg.

Second was Soriç (1989), who argued forcefully against the use of uncontrolled single-hypothesis testing when many are tested, and used the expected number of false discoveries divided by the number of discoveries as a warning that ‘a large part of statistical discoveries may be wrong’. Reading Soriç (1989) we realized that with V being the number of type I errors made, out of the R rejected, by defining $\text{FDR}_{-1}=E(V)/E(R)$ (Fdr in Efron (2008) and earlier) we obtain a very appealing error rate, that rather than being merely a warning can serve as a worthy goal to control. Moreover, considering these quantities as a function of the level α at which the individual testing is done, a plausible estimator for the FDR is $Q(\alpha)=\alpha m_0/R(\alpha)$. Indeed, the value depends on m_0, but we already had a way to estimate m_0 from our previous work! To our delight, to obtain $\max\{\alpha|Q(\alpha)\leqslant q\}$ we could use a step-up method on the sorted series of p-values (theorem 2 in Benjamini and Hochberg (1995)). In November 1989 we submitted a paper named ‘A synthesis of new approaches to multiple comparisons’.

Readers of the submitted manuscript were concerned with the definition of FDR when $R=0$. We then considered other possible definitions: $\text{FDR}=E(V/R)$, where $V/R=0$ when $R=0$, and $\text{FDR}_{+1}=E(V/R|R>0)$ (pFDR in Storey (2002)). We adopted the FDR because controlling it assured weak control of the familywise error rate $\text{FWER}=\Pr(V\geqslant 1)$ when all hypotheses are true—a property that we considered essential for use in medical research, and a property that the other two definitions could not enjoy.

Address for correspondence: Yoav Benjamini, Schreiber School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, 69978 Tel Aviv, Israel.
E-mail: ybenja@tau.ac.il

1369–7412/10/72405

5 years and three journals later the paper was accepted for publication. Along the way, one of the many attempts to prove an FDR property by induction, making use of m as an upper bound for m_0, turned into success, so we took out of the paper the adaptive stage of estimating m_0, merely noting this power increasing possibility. It was therefore no longer a synthesis of new approaches, so the name was changed accordingly.

2. Success of the false discovery rate idea

Acceptance of the FDR idea remained slow even after Benjamini and Hochberg (1995) was published. Our original paper, with the estimated m_0, appeared only 5 years and two journals later (Benjamini and Hochberg, 2000). Other papers on the FDR by various researchers encountered similar difficulties. The dramatic change in attitude came when genetic research took a new dimension, in quantitative trait loci and microarrays analyses, where the number of hypotheses tested in an experiment reached thousands. This seemed unthinkable 10 years earlier: for example, our simulations in Benjamini and Hochberg (1995) had been criticized for considering 4–64 hypotheses, as 'no one uses multiple comparisons for problems with 50 or 100 tested hypotheses'. Alas, facing the new challenges, tools that balance multiplicity control and power were needed, and FDR methodology could yield useful answers.

By the year 2000 quite a few groups of statisticians were working in the area, and results in FDR theory, methods and applications started to flow. It became clear that the FDR is a very intuitive concept, which adaptively spans the entire range from extreme multiplicity control to none, depending on the data encountered. If the data justify—even in very large problems—it is very permissive. In sparse problem it acts close to control of FWER; still the extra allowance gives it an edge in performance (see Section 3.4). Finally, it is interpretable from different points of view: frequentist, Bayesian, empirical Bayes and decision theory.

A full review of the FDR developments is beyond the scope of this short note. Doing injustice to many, I follow the guidelines of the editors and outline progress and conceptual developments in FDR research related to my own involvement.

3. Later progress

3.1. Estimation of m_0 (or $p_0 = m_0/m$)

Much research effort has gone into developing new estimation methods. Storey (2002) suggested the use of $\#(p\text{-values} > c)/(1-c)$. In Storey and Tibshirani (2003) the cut-off point c has been chosen via bootstrapping. Fitting different mixture models to the distribution of the test statistics, to their p-values, or their transformed z-values, and estimating the proportion of distribution that is attributed to values under the true null hypotheses are some directions taken, either parametric (e.g. by Allison and co-workers), or non-parametrically (by Genovese and Wasserman (2002) and Efron and co-workers).

It should be emphasized that, once an estimator $\mathbf{m}_0$ of m_0 is inserted into the procedure in Benjamini and Hochberg (1995), it is no longer guaranteed to achieve FDR control at the desired level. Adjustments may be needed in the estimator: see Storey *et al.* (2004), Benjamini *et al.* (2006), Gavrilov *et al.* (2009) and Blanchard and Roquain (2009) for modifications introduced to natural estimators. Such adjustments are especially crucial under dependence and when $m_0/m \sim 1$ (including $m_0/m = 1$).

3.2. Addressing dependence

Independence of test statistics was assumed in Benjamini and Hochberg (1995). Addressing positive dependence by Benjamini and Yekutieli (2001) was essential in assuring users that the simple procedure in Benjamini and Hochberg (1995) was safe to use in many situations arising in practice. It built on the work of Sarkar (1998) and was followed by the work of Sarkar, Finner and co-workers. The modification to general dependence is often not needed: convincing simutheoretical evidence indicates that the same holds for two-sided z-tests with any correlation structure (Reiner-Benaim, 2007), but the theory awaits a complete proof. Other theoretical puzzles remain open, e.g. the pairwise comparisons setting. It is well documented that the FDR is less than 0.05, but by how much? On establishing theoretical results, a better procedure can be designed.

The other way to address dependence is by bootstrapping and rerandomization. See Yekutieli and Benjamini (1999), Storey and Tibshirani (2003) and van der Laan and Dudoit (2007).

3.3. *Bayes and empirical Bayes approaches to false discivery rate*

Much research has been devoted to FDR ideas from the Bayes and the empirical Bayes perspectives, and use the insight thus gained to derive new theory and methodologies. The empirical Bayes approach to the FDR has been reviewed by Efron (2008), in a very nice and accessible way. The purely Bayesian work merits a separate review.

4. Conceptual developments

4.1. *'Testimation'*

In Abramovich and Benjamini (1996) we first suggested the use of the procedure in Benjamini and Hochberg (1995) for thresholding wavelets coefficients when denoising signals. Abramovich *et al.* (2006) (also Johnstone's Wald Lecture at the meeting of the Institute of Mathematical Statistics in Baltimore in 1999) showed that sparse signals could be retrieved by this method at the optimal rate and with the correct constant. Moreover, optimality holds for both signals that are sparse because most coefficients are 0, and when none is 0 but their ordered size decays fast. That testing approach deals well and the latter justifies the importance of simple null hypotheses in testing, even if a null hypothesis is never exactly true in practice: estimation following selection by testing, or 'testimation', is a practice that can be justified by theory.

This work progressed in different directions by different researchers:

(a) increasing its generality (Donoho and Jin, 2006);
(b) showing that the FDR can be an effective model selection criterion, as it can be translated into a penalty function (Benjamini and Gavrilov, 2009);
(c) investigating the theoretical boundaries of detection of signals in the space that is spanned by their sparsity and signal-to-noise ratio, using different multiple-testing approaches, FWER, the FDR and their 'higher criticism' that controls the FWER in the weak sense (Donoho and Jin, 2004).

4.2. *False coverage statement rate and selective inference*

In Benjamini and Yekutieli (2005) we demonstrated that confidence intervals that are constructed for selected parameters only, where selection depends on the observed values, cannot ensure nominal coverage even on average. The false coverage statement rate FCR was offered as a criterion that parallels the FDR, except that a discovery is replaced by 'a confidence statement on a selected parameter is made' and a false discovery is replaced by 'a confidence statement on a selected parameter fails to cover the parameter'. A general procedure was given, where marginal $1-q|S|/m$ confidence intervals are constructed for the $|S|$ selected parameters, where S is the (data-dependent) selected set.

Working on FCR revealed to us that simultaneous and selective inference, the traditional justifications for multiple-comparisons procedures, are two distinct goals. This fact was masked because all FWER controlling methods offer simultaneous inference, which in turn implies selective inference. However, selective inference can be a goal by itself. We now view the FDR and FCR as concepts that address directly the dangers that are caused by selective inference—the reporting, or highlighting, or attending only to the significant findings, which may alter the meaning of the reported p-values and confidence intervals—while giving up simultaneous inference. In most large problems we care only about the effect of selection.

Efron (2008) has discussed the issue, and the associated difficulties, from the empirical Bayes approach. Yekutieli (2010) has discussed it from both the Bayesian and the empirical Bayes approach and addressed formally the effects of selection. Benjamini *et al.* (2009) have drawn attention to this problem in replicability studies of genomewise scans for association with a disease, where detailed inference is made on fewer than a dozen locations among the 400000 scanned in four studies.

4.3. Multiple multiplicity error rates

Once the FDR managed to break the dichotomy of 'don't worry be happy' unadjusted approach, *versus* the 'panic' FWER approach, many other error rates that try to take some middle way were offered. A partial list includes the false exceedance rate $\Pr(V/R>q)$, k-FDR and k-FWER, and the local FDR. See Benjamini (2010) for references and discussion.

This multiplicity of multiplicity error rates should be welcomed. Each of them might be appropriate and useful for some inferential situation: personal decision making, policy decision making, monitoring, scientific communication or licensing. Alas, so far practitioners have been offered little advice about which error rate each situation requires, which is a condition that we should change. Otherwise, users will again avoid using control of error rates in a meaningful way and rely instead on *ad hoc* solutions (say the use of 10^{-5} as a threshold of significance in some genomewise association scans).

Note that the estimation of error rates is sometimes presented as a very different approach to controlling error rates. I view this as a nuance rather than distinction, as one should care about the properties of a procedure that selects a set of discoveries according to the value of the estimator.

5. Conclusion

It was no surprise to me that practitioners have embraced the FDR approach. I also expected the theoretical developments that would be needed to expand the set of relevant tools (addressing dependence, other estimation methods for m_0 and model selection). I am surprised (and delighted) that the FDR idea generated such diverse interest from the theoretical point of view, extending and expanding our knowledge in so many directions. A remaining puzzle for me is whether FDR control is a manifestation of another principle, be it empirical Bayes, decision theoretic, minimum description length and such, or whether it is a principle of its own, which in some setting coincides with another principle.

Ending with a personal note: I am extremely sad that Yosi Hochberg could not join me in writing this review, nor travel to Edinburgh to enjoy the 'retrospectively read paper' event, because of his health problems in recent years. I am sure that the readers will join me in wishing him good health.

Acknowledgements

An important motivation for this note is to be able to acknowledge the support I had received from friends and colleagues at early stages of our work. Thanks go to Julie Shaffer, Henry Braun, Larry Brown, Peter Westfall and Mark Low. Thanks also go to my many co-authors who invested their time and efforts in FDR research before it was well known or well accepted: Felix Abramovich, Dave Donoho, Iain Johnstone, Abba Krieger and Dani Yekutieli. Thanks go to Mike Titterington for his editorial advice during the preparation of the manuscript. Finally, I would like to cherish the memory of John Tukey, whose encouragement and advice had escorted me since the first draft, and that of Erich Lehmann, whose support was more subtle but important.

References

Abramovich, F. and Benjamini, Y. (1996) Adaptive thresholding of wavelet coefficients. *Computnl Statist. Data Anal.*, **22**, 351–361.

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. M. (2006) Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34**, 584–653.

Benjamini, Y. (2010) Simultaneous and selective inference: current successes and future challenges. *Biometr. J.*, to be published.

Benjamini, Y. and Gavrilov, Y. (2009) A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Statist.*, **3**, 179–198.

Benjamini, Y., Heller, R. and Yekutieli, Y. (2009) Selective inference in complex research. *Phil. Trans. R. Soc. Lond.* A, **367**, 4255–4271.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.

Benjamini, Y., Krieger, M. A. and Yekutieli, D. (2006) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93**, 491–507.

Benjamini, Y. and Yekutieli, Y. (2001) The control of the False Discovery Rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Benjamini, Y. and Yekutieli, Y. (2005) False discovery rate controlling confidence intervals for selected parameters. *J. Am. Statist. Ass.*, **100**, 71–80.

Blanchard, G. and Roquain, E. (2009) Adaptive FDR control under independence and dependence. *J. Mach. Learn. Res.*, **10**, 2837–2871.

Donoho, D. and Jin, J. S. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

Donoho, D. and Jin, J. (2006) Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.*, **34**, 2980–3018.

Efron, B. (2008) Microarrays, empirical Bayes and the two groups model. *Statist. Sci.*, **23**, 1–22.

Gavrilov, Y., Benjamini, Y. and Sarkar, S. (2009) An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.*, **37**, 619–629.

Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B, **64**, 499–517.

Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Statist. Med.*, **9**, 811–818.

van der Laan, M. J. and Dudoit, S. (2007) *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Reiner-Benaim, A. (2007) FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometr. J.*, **49**, 107–126.

Sarkar, S. K. (1998) Some probability inequalities for ordered MTP2 random variables: a proof of Sime's conjecture. *Ann. Statist.*, **26**, 494–504.

Schweder, T. and Spjøtvoll, E. (1982) Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.

Soriç, B. (1989) Statistical "discoveries" and effect size estimation. *J. Am. Statist. Ass.*, **84**, 608–610.

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

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.
Storey, J. D. and Tibshirani, R. J. (2003) Statistical significance for genome-wide experiments. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.
Yekutieli, D. (2010) Adjusted Bayesian inference for selected parameters. *Preprint arXiv: 0801.0499v4*.
Yekutieli, D. and Benjamini, Y. (1999) Resampling based False Discovery Rate controlling procedure for dependent test statistics. *J. Statist. Planng Inf.*, **82**, 171–196.

Comments on the presentation

Ernst Wit (*University of Groningen*)
When I was Secretary of the Society's Research Section, it was proposed in the Section to extend the existing concept of 'read' papers, in which important novel research is presented, to 'retrospectively read' papers, older papers that have been published in the Society's journals, which subsequently have become very influential. From the very start the Research Section Committee faced several problems: would the original paper be read or would the authors be asked for an 'update'? Would it become a *Festschrift*? The Committee wisely decided to require the paper to be in an area that was still actively under development, allowing the authors to provide a background to the original paper and guaranteeing that the discussion would involve recent research contributions.

The only problematic aspect of this whole meeting, perhaps, is the role of the proposer of the vote of thanks: in fact, Benjamini's contribution above is a clear analysis and celebration of the original paper and so what is left for me to do? Benjamini has identified many of the strengths of the original paper and the way that they have been developed in the years thereafter; he has shown several open problems, both theoretic ones and related to particular applications. And the fact that this paper has been chosen as a re-read paper is in itself the largest vote of thanks, as we have the advantage of hindsight this time around. Nevertheless, I would like to add some words to Benjamini's account.

Multiple testing
Hypothesis testing is a difficult idea. We all know it when we try to explain the concept in a statistics service course. Students see it as formulaic. In fact, it *is* formulaic. Suddenly they are asked to care about the probability that such a *kind* of data could occur if a particular statement about the parameters was in fact the case. Multiple testing does not make it any easier. p-value *corrections*—correcting what?—make them lose sight of what it is that they are actually calculating. According to me, its most important contribution was the way that Benjamini and Hochberg (1995) opened up the field of hypothesis testing in general, and multiple testing in particular. The title of the paper testifies to clear foresight: it is exactly a *practical* approach, suited to practitioners dealing with multiple questions.

However, it is still early days and not infrequently papers citing Benjamini and Hochberg (1995) are stuck in their old ways. For example, Patti *et al.* (2003) (which is the ninth most highly cited paper that cites Benjamini and Hochberg (1995)) when testing 7129 gene sequences wrote

> 'No single gene remained differentially expressed after Benjamini-Hochberg multiple comparison testing'.

Other papers talk about Benjamini–Hochberg 'corrections' or about 'significance' (Weisberg *et al.* (2003), which is the third most highly cited paper that cites Benjamini and Hochberg (1995)). However, there are no corrections and there is no significance. Benjamini and Hochberg (1995) introduced two things: a new error rate and an algorithm that under certain conditions controls that error rate. Old ways may die slowly, but Benjamini and Hochberg (1995) gave us an intuitive way to understand multiple testing.

Interpretation
Despite its apparent simplicity, the interpretation is more complicated than we might expect at first sight. If we control the false discovery rate FDR at level α and reject r hypotheses, it does not mean that αr are wrongly rejected, nor even that *on average* αr are wrongly rejected. It means that this procedure when applied many times to this problem on average rejects a fraction α (or less) incorrectly. If we know that the number of rejections is r, we could potentially calculate $E(V/r)$, a conditional FDR, which might be different from α.

There is also the confusion about estimation and control. The Benjamini–Hochberg method does not estimate FDR. In fact, it does things the other way around. It selects the level α and adopts a rejection procedure such that FDR $\leqslant \alpha$. However, later work has blurred this distinction: there are methods that given a rejection procedure $\mathcal{R}$ estimate the unknown quantity $E_{\mathcal{R}} V/R$.

Impact

In many ways, Benjamini and Hochberg (1995) is a very successful paper. Its influence is clear from its 4967 citations (according to the *Web of Science* at the time of this session), which are still on the rise each year as can be seen in Fig. 1. Although 607 of these are in the area of statistics and probability, the majority of these publications are in the life sciences, from genetics to biochemistry, from oncology to plant sciences, reflecting in large part the use of FDR in microarray-related research. Importantly, citations in other high dimensional application areas, such as neural imaging, are on the rise also, showing its ability to be applied in many diverse types of application. The list of the 10 highest cited papers that cite Benjamini and Hochberg (1995), which is shown in Table 1, is particularly interesting, because it includes six statistical papers, suggesting that further theoretical and methodological developments of the method have had significant influence.

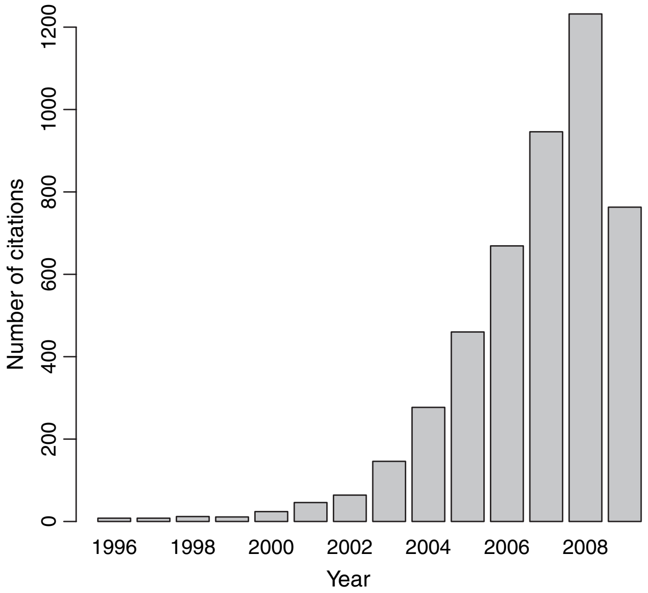

Fig. 1. Rapidly increasing number of citations of Benjamini and Hochberg (1995), suggesting that its influence has not yet reached its peak (note that the figure for 2009 is only partially shown)

Table 1. 10 most cited papers that cite Benjamini and Hochberg (1995)

Rank	*Article citing Benjamini and Hochberg (1995)*	*Number of citations*
1	Tusher *et al.* (2001)	3723
2	Storey and Tibshirani (2003)	1412
3	Weisberg *et al.* (2003)	1187
4	Genovese *et al.* (2002)	1020
5	Storey (2002)	726
6	Wilkinson (1999)	652
7	Benjamini and Yekutieli (2001)	584
8	Wacholder *et al.* (2004)	486
9	Patti *et al.* (2003)	479
10	Dudoit *et al.* (2002)	459

Summing up, Benjamini and Hochberg (1995) introduced an important new concept, the false discovery rate, and presented a practical approach to implement this concept. Both had enormous practical impact and at the same time the methodology stood at the beginning of a line of active statistical research, something that we would normally require from a read paper. It is therefore extremely appropriate that the Society's Research Section Committee has selected Benjamini and Hochberg (1995) as the first retrospectively read paper and I am happy to propose the vote of thanks.

V. T. Farewell (*Medical Research Council Biostatistics Unit, Cambridge*)
Accepting the 1932 Democratic nomination for President, Franklin D. Roosevelt said: 'I pledge you, I pledge myself, to a new deal for the American people'. His new deal was the real deal for economic recovery. We are reflecting on the false discovery rate (FDR) and, 14 years after Benjamini and Hochberg's (1995) paper on this topic, the 'new deal' offered by this 'FDR' has also proved to be a 'real deal'.

So, first, let me thank Professor Benjamini, both for his 1995 paper and for his interesting retrospective look at the paper, its genesis and its impact.

In his presentation, Professor Benjamini indicated that publication delays mentioned arose because reviews of the initial, and subsequent, FDR papers were split. For one particular paper of my own which received similar split reviews, the editor reported this divergence of views and indicated that, for this reason, the journal would publish the paper. In my editorial experience, split reviews may arise because of genuine scientific disagreements, which I assume was the editor's assessment in this particular case. In another scenario, one reviewer may be better equipped to evaluate the paper. And, finally, divergence may arise when the paper is, explicitly or implicitly, challenging an established viewpoint and therefore may elicit a defensive reaction from those with an investment in the established viewpoint, but not necessarily from others.

What would have been my reaction if asked to review the submitted FDR paper?

Multiple-comparison procedures can be discussed in terms of questions of interest. Cox (1965) reflected this and pointed out that probabilities regarding the simultaneous correctness of many statements may not always be of direct relevance, particularly when there is interest in a specific statement. In the context of clinical trials, Richard Cook and I elaborated this viewpoint, expressed by Cox in one and a half pages, in our paper of 18 pages (Cook and Farewell, 1996), under consideration around the same time as the Benjamini and Hochberg submission.

For illustration, consider a current study of cognitive function in various diseased populations. With a subset of the study data, the results from a battery of seven neuropsychological tests can be compared in multiple sclerosis and systemic *lupus erythematosus* patients. With adjustment for age and education, the levels of significance associated with the seven tests are 0.036, 0.039, 0.075, 0.284, 0.510 and 0.813. With a 5% FDR, no comparison would be deemed significant. However, primary interest is in overall performance for the battery of tests and, for example, O'Brien's generalized least squares rank-based procedure (O'Brien, 1984) generates a global level of significance of 0.013. Here, an FDR procedure may not be answering the most useful question.

I would also have commented on the example that was used by Benjamini and Hochberg (1995). The original clinical paper examined data on 15 cardiac and other events, one of which was mortality. A Bonferroni approach with a 5% familywise error rate FWER would therefore not have rejected the null hypothesis of no mortality difference, which had a nominal level of significance of 0.0095. In contrast, a 5% FDR procedure would have rejected this hypothesis, and three other hypotheses also rejected by the Bonferroni procedure. But, surely, mortality is a clinical outcome of very specific importance, perhaps the archetypal 'primary outcome'. Should our inference concerning this depend on conclusions about other aspects of the treatment comparison? I would have disagreed with the authors that only with an FDR procedure is there 'appropriate confidence' to support a difference in mortality.

Nevertheless, when compared with procedures which control FWER, the probability of one or more false positive tests, the FDR addresses a very different type of question. The focus is on an 'error rate' that is defined in terms of the 'rejected' hypotheses, functioning in some respect like a positive predictive value. In a situation when probability statements concerning simultaneous correctness of multiple statements may be useful, the FDR approach should surely have been seen as at least equally worthy of consideration as the FWER-approach, potentially having advantages in some circumstances. Thus, I should like to think that I would have supported publication of the paper, if certain *caveats* had been made and the example was seen as illustrative and not definitive. I wonder therefore whether there was, in fact, either a respectable difference of scientific opinion or an aspect of defensiveness to those early negative reviews. If so, it might serve as a useful caution when writing reviews.

Subsequently, as Professor Benjamini has indicated, the FDR procedure came into its own with the new challenges of genetic and genomic research when a very large number of simultaneous significance tests were being performed. Other areas of application have also emerged, the comparison of many healthcare providers (Jones *et al.*, 2009) being a notable illustration. Questions for which the FDR was suited have arrived! Although there may still be some concerns about procedures based solely on levels of significance and about the exact nature of the probabilities behind the procedures, in these applications, the FDR has, minimally, practical usefulness. This was characterized by Cox (1965) as 'giving a conservative bound for the effect of selection, rather than in giving an "exact" solution'.

It gives me great pleasure, therefore, to second the vote of thanks to Professor Benjamini for his important 1995 paper and for his retrospective look at it.

The vote of thanks was passed by acclamation.

José A. Ferreira (*National Institute for Public Health and the Environment, Bilthoven*)
I have found it very interesting to read Benjamini's views on his method and to learn about the early adventures and tribulations of the false discovery rate (FDR). I shall discuss two aspects of the FDR which Benjamini did not emphasize in his presentation and which are relevant when m is large:

(a) the Benjamini–Hochberg (BH) method works more generally and often in a stronger sense than originally thought, and in a sense it is optimal;
(b) however, the few assumptions that are required by the method should be checked as much as possible, and they call for the development of 'diagnostic' procedures.

Regarding (a), suppose that of the m statistics $T_1, \ldots, T_m$ the first m_0 are computed 'under the null hypothesis' and have the same distribution function (DF) F whereas the other $m - m_0$ tend to take *smaller* values; for example, the statistics could be p-values and F the uniform DF. By rejecting all hypotheses whose statistics fall strictly below the 'threshold' t we incur a false discovery *proportion* FDP of

$$\mathrm{FDP}_m(t) := \frac{\sum_{i=1}^{m_0} \mathbf{1}_{\{T_i<t\}}}{\sum_{i=1}^{m} \mathbf{1}_{\{T_i<t\}}} \equiv \frac{m_0}{m}\,\frac{F_{m_0}(t-)}{H_m(t-)} \equiv \pi_m \frac{F_{m_0}(t-)}{H_m(t-)},$$

where H_m is the empirical DF of the T_is, F_{m_0} that of $T_1, \ldots, T_{m_0}$ and $\pi_m = m_0/m$. This provided $H_m(t-) > 0$, for otherwise no hypotheses are rejected and $\mathrm{FDP}_m(t) = 0$. Consequently, by rejecting all hypotheses with statistics strictly below the threshold

$$t_m = \sup\{t : \pi_m \, F_{m_0}(t)/H_m(t) \leqslant q\}$$

we can keep FDP $\leqslant q$. Moreover, since we cannot increase t_m without risking an increase in FDP beyond q, and since the bigger the threshold the bigger the number of true discoveries, t_m is optimal. Unfortunately, F_{m_0} being unobservable and π_m unknown, this ideal threshold cannot be determined. But if F_{m_0} is close to F and π_m is not much smaller than a given π_m^* then

$$t_m^* = \sup\{t : \pi_m^* F(t)/H_m(t) \leqslant q\}$$

is close to the ideal threshold t_m and hence controls FDP *approximately* at q and is *close to being optimal.*

As theorem 2 of Benjamini and Hochberg (1995) shows, the procedure based on the approximate threshold t_m^* *is* the BH method, and what the above formulation makes apparent is that, except in 'pathological situations', the method controls FDP in the various senses of convergence (as $m \to \infty$) and under a very wide range of probability models (essentially under any model for which $F_{m_0} \to F$ uniformly). Of course, just as with most limit theorems in statistics, the extent to which the 'BH limit theorem' applies to a given data set is unascertainable, but that does not make it less useful.

Regarding (b), if the T_is are rank statistics (say) the method may work accurately even if the F that is used to define t_m^* is the *asymptotic* DF of the T_is. More generally, F_{m_0} is often closer to some $\tilde{F}$ than to F but the resulting 'disturbance' on the FDR is small because $\tilde{F}$ is itself close to F. However, sometimes the postulated F deviates grossly from the correct DF in the tail(s), leading to a serious underestimation or overestimation of the FDR; I suspect that this is often so when the T_is are p-values computed from t-statistics, and perhaps even more so when they are based on regression models. This calls for the development of methods (mainly graphical and necessarily informal, I think) for checking the 'approximate validity' of the BH theorem.

G. Green and P. J. Diggle (*Lancaster University*)

The original paper by Benjamini and Hochberg (1995) has achieved well-deserved fame through its development of the false discovery rate (FDR) as an alternative to classical methods of adjustment for multiple significance testing, which fail to provide reasonable procedures when the number of tests is large. The concept of FDR control has been embraced by the genomics community, where a typical objective of statistical analysis is to identify differentially expressed genes. Nevertheless, we still see papers in non-statistical journals that use the naive procedure of testing at extreme levels of significance, say $p = 10^{-5}$ or $p = 10^{-6}$. The Society's decision to highlight the importance of the 1995 paper, and subsequent work by its authors, is therefore very welcome.

The problem of detecting differentially expressed genes can be cast, apparently quite naturally, in a classical hypothesis testing framework, formulating the null hypothesis as $H_{0,g}$: gene g is not differentially expressed, against the two-sided alternative for $g = 1, \ldots, G$. Typically G is of the order of thousands and the procedure of Benjamini and Hochberg (1995) provides a pragmatic response to the attendant multiple-testing problem.

The detection of differential expression can be considered as a special case of gene profiling where, more generally, the aim is to identify genes which show compelling evidence in favour of satisfying some predefined criterion or profile. In our opinion, although classical testing addresses the common aim of detecting differential expression, it fails to extend adequately to many less standard questions of gene profiling. For example, in the context of a microarray time course experiment, Tuke *et al.* (2008) sought to identify genes that show, simultaneously, evidence in favour of differential expression across one pair of time points and equal (i.e. non-differential) expression across another pair of time points. Using the equivalence testing paradigm (see Wellek (2002)), Tuke *et al.* (2008) developed an approach based on the intersection–union test principle (Berger, 1982), resulting in a conservative test. We are unaware of an exact test which responds to such a question.

We would argue that questions of this kind are better addressed as prediction problems, rather than through any form of hypothesis testing. Green (2008) has developed the following model-based approach for identifying genes that satisfy some predefined profile. Let Y_{gtr} be a suitably preprocessed, log-scale-measured gene expression for gene g, treatment t and replicate r, $g = 1, \ldots, G$, $t = 1, \ldots, T$, $r = 1, \ldots, R$. We suppose that $Y_{gtr} = \mu_g + W_{gt} + Z_{gtr}$, specifying a random gene–treatment interaction effect; typically, the main effect parameters μ_g may be corrupted by the preprocessing and, in any event, are of limited scientific interest. Fitting the model by using either restricted maximum likelihood or Bayesian inference enables evaluation of the joint predictive distribution $[\mathbf{W}_g|\mathbf{Y}]$ for all $g = 1, \ldots, G$ and, more interestingly, of predictive probabilities which reflect the degree to which a gene satisfies some predefined profile. For example, the probability $P(|W_{gt} - W_{gt'}| > d|\mathbf{Y})$ reflects the evidence in favour of gene g being differentially expressed across treatments t and t'. Similarly, the probability $P(\{|W_{gt} - W_{gt'}| > d\} \cap \{|W_{gt''} - W_{gt'''}| < d\}|\mathbf{Y})$ reflects the evidence in favour of gene g being, simultaneously, differentially expressed across t and t', but equally expressed across t'' and t'''. In its simplest form, the model assumes normally distributed random effects, but we have also developed a hierarchical extension in which the variances of the W_{gt} are themselves given scaled inverse χ^2-distributions, to capture non-Gaussian behaviour that, in our experience, is typical of differences in measured log-expression levels.

Author's response

I thank the discussants for their votes of thanks and compliments, and their heartwarming words during their oral presentation, as well as for their illuminating remarks.

Professor Ferreira commented on the need to verify the assumptions underlying the use of the false discovery rate controlling procedure in Benjamini and Hochberg (1995). Interestingly, the method works even when the asymptotic assumptions are clearly violated, as in many-to-one comparisons. Still, the appropriateness relies on valid (conservative) p-values. It is therefore valid for the rank statistics mentioned, but their discrete nature offers potential for improvement.

Professor Farewell gave an example comparing two groups via seven neuropsychological tests, demonstrating that a combination test is more powerful than the procedure in Benjamini and Hochberg (1995). I argued that an important and often neglected issue in practice is a judicious choice of the error rate, and this example is a case in point. If researchers indeed treat the seven tests as a single battery—their implied concern is about a weak familywise error rate (under the global null hypothesis), and the most powerful method assuring it is appropriate. Only if they would further wish to conclude that the difference hinges on some specific tests should they care about the false discovery rate, familywise error rate or their likes, and act accordingly.

Dr Green prefers predictive probabilities that reflect the degree to which a gene satisfies some predefined profile to p-values based on testing. I wonder how much worse did screening over the conjunction hypothesis for each gene perform? More importantly, I would argue that screening across all genes, and selecting the few with 'high predictive probabilities' (low prediction error probabilities) should be accompanied by a procedure addressing the dangers of selective inference.

Professor Wit expresses subtly and politely his dismay with tests and significance statements. I am very much in favour of significance testing in the Fisherian sense, as a formal means to weed out results due to chance variation. I think that the 'testimation' that is discussed in the main paper sheds light from a current point of view on the importance of statistical significance testing.

Both Professor Farewell and Professor Wit emphasized the practical aspect of the false discovery rate approach. I agree, and I further believe that the practical needs will continue to motivate much of the future development. In brain imaging research, for example, involving of the order of 50000 voxels per analysis, the goals of the inference are still debated: should researchers care about the proportion of false discoveries among voxels found active, or the proportion of errors among active regions (contiguous sets of voxels)? Maybe neither is of interest, but rather are topological features such as peaks and cusps? Monitoring healthcare, mentioned by Professor Farewell, is another area where the choices of units of interest need not be the obvious ones. Research areas that require conceptual, theoretical and methodological developments are bound to appear.

The increasing size of the problems attended in practice, contrasted with the few potential discoveries, will continue to challenge researchers. I see three promising directions. *Pooling hypotheses to clusters* is often a way to gain power. In the testing for active voxels in brain imaging experiments, clusters can be based on a moving window (Pacificoa *et al.*, 2007), or a pilot study (Benjamini and Heller, 2007). The clusters need not be of the same size and shape, are of neurological relevance and are tested by using a combining statistic for each cluster. We gain power both from combining the evidence over clusters, and from the smaller number of tests.

Testing for active voxels within the clusters that are found active is a natural way to continue (Benjamini and Heller, 2007). This is an example of *hierarchical testing* of a tree of hypotheses, where a subfamily of a branch is tested only after the node from which it branches has been tested and rejected. The general tree structure can be of varying depth and size, and opportunities for power gain arise when hypotheses in a branch tend to be true or false together. Reiner-Benaim *et al.* (2007) have presented a framework, and Yekutieli (2008) has provided theoretical analysis for simple but important cases. Zehetmayer *et al.* (2005) used a hierarchical approach for screening experiments, and Meinshausen (2008) for testing the importance of variables in a regression model (from the familywise error rate perspective). The general hierarchical approach is extremely flexible and promising, yet it awaits future developments.

Employing weights to differentiate between the hypotheses tested is another direction. The weights may incorporate differing importance of the hypotheses (Benjamini and Hochberg, 1997), or different prospects for showing effects (Genovese *et al.*, 2006). The weights can be based on outside information, or on information from initial testing or on the size of the set (Benjamini and Heller, 2007). There are many conceivable variations of combination of all three directions, as Hu *et al.* (2009) have demonstrated.

Of course directions that we are currently unaware of may end up being most fruitful.

References in the comments

Benjamini, Y. and Heller, R. (2007) False discovery rates for spatial signals. *J. Am. Statist. Ass.*, **102**, 1272–1281.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (1997) Multiple hypothesis testing with weights. *Scand. J. Statist.*, **24**, 407–418.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Berger, R. L. (1982) Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, **24**, 295–300.

Cook, R. J. and Farewell, V. T. (1996) Multiplicity considerations in the design and analysis of clinical trials. *J. R. Statist. Soc.* A, **159**, 93–110.

Cox, D. R. (1965) A remark on multiple comparison methods. *Technometrics*, **7**, 223–224.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002) Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statist. Sin.*, **12**, 111–139.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002) Thresholding of statistical maps in functional neuroimaging using false discovery rate. *Neuroimage*, **15**, 870–878.

Genovese, C. R., Roeder, K. and Wasserman, L. (2006) False discovery control with *P*-value weighting. *Biometrika*, **93**, 509–524.

Green, G. (2008) Statistical methods for the analysis of microarray data. *PhD Thesis*. Lancaster University, Lancaster.

Hu, J. X., Zhao, H. and Zhou, H. H. (2009) Multiple hypothesis testing with groups. *Technical Report*. Yale University, New Haven.

Jones, H. E., Ohlssen, D. I. and Spiegelhalter, D. J. (2008) Control of the false discovery rate accounts for multiple testing in comparisons of healthcare providers. *J. Clin. Epidem.*, **61**, 232–240.

Meinshausen, N. (2008) Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278.

O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.

Pacificoa, M. P., Genovese, C., Verdinelli, J. and Wasserman, L. (2007) Scan clustering: a false discovery approach. *J. Multiv. Anal.*, **98**, 1441–1469.

Patti, M. E., Butte, A. J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., Landaker, E. J., Goldfine, A. B., Mun, E., DeFronzo, R., Finlayson, J., Kahn, C. R. and Mandarino, L. J. (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: potential role of pgc1and nrf1. *Proc. Natn. Acad. Sci. USA*, **100**, 8466–8471.

Reiner-Benaim, A., Yekutieli, D., Letwin, N. E., Elmer, G. I., Lee, N. H., Kafkafi, N. and Benjamini, Y. (2007) Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay. *Bioinformatics*, **23**, 2239–2246.

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

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.

Tuke, J., Glonek, G. F. V. and Solomon, P. J. (2008) Gene profiling for determining pluripotent genes in a time course microarray experiment. *Biostatistics*, **10**, 80–93.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natn. Acad. Sci. USA*, **98**, 5116–5121.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormi, L. and Rothman, N. (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natn. Cancer Inst.*, **96**, 434–442.

Weisberg, S., McCann, D., Desai, M., Rosenbaum, M., Leibel, R. L. and Ferrante, A. W. (2003) Obesity is associated with macrophage accumulation in adipose tissue. *J. Clin. Investgn*, **112**, 1796–1808.

Wellek, S. (2002) *Testing Statistical Hypotheses of Equivalence*. New York: Chapman and Hall.

Wilkinson, L. (1999) Statistical methods in psychology journals—guidelines and explanations. *Am. Psychol.*, **54**, 594–604.

Yekutieli, D. (2008) Hierarchical false discovery rate controlling methodology. *J. Am. Statist. Ass.*, **103**, 309–316.

Zehetmayer, S., Bauer, P. and Posch, M. (2005) Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, **21**, 3771–3777.