

Whittemoretxt050806.tex

A Bayesian False Discovery Rate for Multiple Testing

Alice S. Whittemore

Department of Health Research and Policy

Stanford University School of Medicine

Correspondence Address: Alice S. Whittemore, Department of Health Research and Policy, HRP Redwood Building, Room T204, Stanford University School of Medicine, Stanford, CA 94305-5405.
Email: alicew@stanford.edu; Phone: 650 723-5460

Acknowledgements: This research was supported by NIH grant number CA94069. The author thanks Raymond R. Balise for help with data analysis.

ABSTRACT *Case-control studies of genetic polymorphisms and gene-environment interactions are reporting large numbers of statistically significant associations, many of which are likely to be spurious. This problem reflects the low prior probability that any one null hypothesis is false, and the large number of test results reported for a given study. In a Bayesian approach to the low prior probabilities, Wacholder et al. (2004) suggest supplementing the p-value for a hypothesis with its posterior probability given the study data. In a frequentist approach to the test multiplicity, Benjamini & Hochberg (1995) propose a hypothesis-rejection rule that provides greater statistical power by controlling the false discovery rate rather than the family-wise error rate controlled by the Bonferroni correction. This paper defines a Bayes false discovery rate and proposes a Bayes-based rejection rule for controlling it. The method, which combines the Bayesian approach of Wacholder et al. with the frequentist approach of Benjamini & Hochberg, is used to evaluate the associations reported in a case-control study of breast cancer risk and genetic polymorphisms of genes involved in the repair of double-strand DNA breaks.*

KEY WORDS: Bayes, breast cancer, false discovery rate, false positive report probability, haplotypes, multiple comparisons, single nucleotide polymorphism

1. **Introduction.** The genetics literature is burgeoning with studies relating disease risk to polymorphisms of specific genes, and studies of interactions between these genes and environmental factors. Often, initial findings are not replicated in subsequent studies. This lack of reproducibility reflects two difficulties. The first is the "poor odds" problem, i.e. the low probability that any statistically significant finding actually is true. The second is the multiplicity problem inherent in the large number of statistical inferences extracted from the study data. When pursuing such multiple inferences, investigators tend to select those with the smallest p-values for emphasis, discussion and support of

conclusions. However an unguarded use of single-inference procedures results in a greatly increased false positive rate. To address this multiplicity problem, classical multiple comparison procedures aim to control the probability of committing any type-I error in the family of tests under simultaneous consideration. This probability is called the family-wise error rate (FWER). However control of the FWER is often purchased at the price of substantial power loss.

Recent work has brought fresh viewpoints to both these problems. For the poor odds problem, Wacholder *et al.* (2004) propose a Bayesian approach that replaces the p-value for a hypothesis with a more informative posterior probability that the null hypothesis is true, obtained by combining its prior probability with its p-value and the study power. For the multiplicity problem, Benjamini & Hochberg (1995) propose controlling not the FWER, but rather a less conservative error measure called the false discovery rate (FDR). The authors provide a rule for controlling the FDR in multiple inferences. But because the rule address the multiplicity problem in a frequentist setting, it does not consider the posterior probabilities proposed by Wacholder *et al.*

This paper defines a Bayesian FDR and provides a simple method for controlling it. Section 2 reviews the frequentist FDR and the rule of Benjamini & Hochberg. Section 3 introduces the Bayes FDR and shows how to control it using the posterior probabilities of Wacholder *et al.* Section 4 illustrates the method with examples, and Section 5 applies it to a study of associations between breast cancer risk and genotypes of genes involved in the repair of double-strand DNA breaks. Section 6 concludes with a discussion of the strengths and limitations of the approach.

2. The test multiplicity problem. Suppose we wish to test m null hypotheses H_i , $i = 1, \dots, m$, using data denoted as Y . Let $X_i = 1$ if H_i is true and $X_i = 0$ if H_i is false, $i = 1, \dots, m$. Let $T_i(Y)$ denote a test statistic for H_i , $i = 1, \dots, m$. The p-value for the hypothesis H_i associated with the observed data $Y = y_{obs}$ is

$$p_i = \Pr(Y \in S_i(y_{obs}) | X_i = 1). \tag{1}$$

Here

$$S_i(y_{obs}) = \{y : T_i(y) \geq T_i(y_{obs})\} \quad (2)$$

is the set of data points y giving test statistics $T_i(y)$ as or more extreme than the observed value $T_i(y_{obs})$. (For definiteness, we have assumed that large values of T_i indicate evidence against H_i , with larger values providing stronger evidence.)

A *rejection rule* is a mapping r from the sample space of Y to a set of vectors $r(y) = (r_1(y), \dots, r_m(y))$, where $r_i(y) = 1$ if H_i is rejected when $Y = y$ and $r_i(y) = 0$ if H_i is accepted. An example of a rejection rule is one that sets $r_i(y) = 1$ if $T_i(y)$ exceeds some predetermined critical value, for $i = 1, \dots, m$. The frequentist FDR associated with rejection rule r is the expected fraction of rejected hypotheses that are true:

$$\varphi(r) = E \left[\frac{\sum_{i=1}^m X_i r_i(Y)}{\sum_{i=1}^m r_i(Y)} \mid X_i, i = 1, \dots, m \right]. \quad (3)$$

In (3) we take $0/0 = 0$. In this definition the expectation is taken with respect to the random variable Y , with the X_i fixed.

Benjamini & Hochberg (1995) proposed a rejection rule r_F designed to control the FDR at a prespecified level α , $0 < \alpha < 1$. To implement it, we order the m labels $i = 1, \dots, m$ according to increasing size of the p-values p_i of (1): $p_1 \leq p_2 \leq \dots \leq p_m$. Then we set

$$r_F = (\underbrace{1, \dots, 1}_k, 0, \dots, 0), \quad (4)$$

where $0 \leq k \leq m$ is the largest value of i for which $p_i \leq \frac{i}{m}\alpha$. The rule r_F rejects all hypotheses whose p-values are at least as small as p_k . Benjamini & Hochberg (1995) proved the following theorem.

Theorem 1. When the test statistics T_1, \dots, T_m are mutually independent, $\varphi(r_F) \leq \alpha$ for all possible values of the vector X .

Benjamini & Yekutieli (2001) proved that Theorem 1 holds under a certain type of stochastic dependence among the test statistics T_i , and Storey & Tibshirani (2003) outlined arguments showing

that the theorem holds asymptotically, as m becomes large, under any form of dependence among the T_i .

3. A Bayesian FDR. In a Bayesian approach to controlling the false positive rate, both the data Y and the vector $X = (X_1, \dots, X_m)$ of indicators for the truth of the m null hypotheses are regarded as random variables. We define the *Bayes false discovery rate* (Bayes FDR) associated with a rejection rule r analogously to the frequentist FDR of (3), but with the expectation taken with respect to the posterior distribution of X given the observed value $r(y_{obs})$ of the rule:

$$\varphi_B(r) = E \left[\frac{\sum_{i=1}^m X_i r_i(Y)}{\sum_{i=1}^m r_i(Y)} \mid r_i(Y) = r_i(y_{obs}), \quad i = 1, \dots, m \right]. \quad (5)$$

Here we again take $0/0 = 0$. Like the frequentist FDR, the Bayes FDR can be interpreted as the expected proportion of incorrect rejections among all rejections of the null hypotheses, with the expectation now taken with respect to the posterior distribution of the hypotheses' truth/falsity, given their acceptance or rejection on the basis of the observed data.

The Bayes FDR has another interpretation, facilitated by rewriting (5) as

$$\varphi_B(r) = \frac{\sum_{i=1}^m \Pr(X_i = 1 \mid r_i(Y) = 1) r_i(y_{obs})}{\sum_{i=1}^m r_i(y_{obs})}. \quad (6)$$

Here $\Pr(X_i = 1 \mid r_i(Y) = 1)$ is the probability that hypothesis H_i is true, given its rejection when rule r is applied to the data Y . This probability has been called the *false positive report probability* (FPRP) by Wacholder *et al.* (2004). According to (6), the Bayes FDR associated with rule r is the mean FPRP, averaged over the rejected hypotheses.

The notion of FPRP suggests a Bayes rejection rule r_B analogous to the frequentist rule r_F . To describe it, we introduce a Bayesian analogue of the p-value p_i , which I call the *b-value*

$$b_i = \Pr(X_i = 1 \mid Y \in S_i(y_{obs})). \quad (7)$$

Expressions (7) and (2) indicate that b_i is the posterior probability that H_i is true, given a test statistic

as extreme as $T_i(y_{obs})$. The b-value b_i depends on the p-value p_i of (1) through the relation

$$b_i = \frac{(1 - \pi_i) p_i}{(1 - \pi_i) p_i + \pi_i q_i}, \quad i = 1, \dots, m. \quad (8)$$

Here π_i is the prior probability that H_i is false, and $q_i = P(Y \in S_i(y_{obs}) | X_i = 0)$ is the probability of obtaining a result as or more extreme than that based on the observed data, when H_i is false. Decision rules based on the b-values have the advantage of accommodating the prior plausibilities of the hypotheses through the π_i , and study power through the q_i . Note from (8) that

$$b_i = (1 + o_i)^{-1}, \quad \text{where} \quad o_i = \frac{\pi_i}{1 - \pi_i} \cdot \frac{q_i}{p_i} \quad (9)$$

is the posterior odds against H_i . Expression (9) shows that o_i is the product of the prior odds $\pi_i / (1 - \pi_i)$ against H_i and the Bayes factor q_i / p_i . Since the b_i are decreasing functions of the o_i , the hypotheses H_i having the smallest b-values (i.e. those most likely to be false) are those with the largest posterior odds against them. Since the b-values depend on more than the p-values, the two rankings $p_1 \leq p_2 \leq \dots \leq p_m$ and $b_1 \leq b_2 \leq \dots \leq b_m$ typically will differ.

The Bayes rule r_B corresponding to a given Bayes FDR α rejects a hypothesis H_i if and only if its b-value $b_i \leq \alpha$. The following analogue of Theorem 1 shows that the rule r_B controls the Bayes FDR.

Theorem 2. $\varphi_B(r_B) \leq \alpha$ for all possible values of the data y_{obs} .

Proof. If none of the m hypotheses are rejected, the Bayes FDR (5) is zero, since we have defined $0/0 = 0$. If $k > 0$ hypotheses are rejected, equation (6) shows that the Bayes FDR is the mean value of their FPRPs. Since the rule requires that all rejected hypotheses have FPRPs at most α , their mean value also must be at most α .

Equation (9) indicates that the Bayes rejection rule r_B rejects a hypotheses when the odds against it are greater than or equal to $(1 - \alpha) / \alpha$. This alternative formulation has heuristic interpretation in applications where odds against a null hypothesis are considered evidence in support of an alternative hypothesis. In this case, the noteworthy study findings are the alternative hypotheses having the largest posterior odds in their favor.

In practice, calculating the b-values requires two types of input. The first consists of the prior probabilities π_1, \dots, π_m . In contrast to applications where flat or uninformative priors are desirable, here we want the value of π_i to reflect both the inherent biological or physical plausibility that H_i is false and the strength and consistency of prior evidence against H_i . Thus the choice of these prior probabilities involves subjective judgement and potential for disagreement (Thomas & Clayton 2004). The second type of input needed are the Bayes factors $q_1/p_1, \dots, q_m/p_m$ of (9), where p_i is the probability of obtaining a test statistic as large as or larger than $T_i(y_{obs})$ when H_i is true, and q_i is the corresponding probability when it is false. Although the p_i are calculated as part of the data analysis, calculating the q_i involves choosing a specific alternative or family of alternatives to H_i , and determining or approximating the tail of the distribution of $T_i(Y)$ under the alternative. The following examples illustrate these issues and choices for the Bayes factor.

4. Examples.

Example 1: independent Gaussian random samples. One of the simplest applications of the approach concerns data $Y = (Y_1, \dots, Y_m)$ where the Y_i are independent variables and each test statistic $T_i(Y)$ depends only on Y_i . For instance, consider m independent random samples $Y_i = Y_{i1}, \dots, Y_{iN_i}$ from Gaussian distributions with means θ_i and known variances σ_i^2 , $i = 1, \dots, m$. We wish to test the m null hypotheses $H_i : \theta_i = 0$ against the alternatives $\theta_i = \theta_i^* > 0$, using the test statistics $T_i(Y) = N_i^{1/2} \bar{Y}_i / \sigma_i$, where \bar{Y}_i is the i^{th} sample mean, $i = 1, \dots, m$. In this example, $p_i = \Phi(-T_i(y_{obs}))$ and $q_i = \Phi[\sqrt{N_i} \theta_i^* / \sigma_i - T_i(y_{obs})]$, where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. The Bayes factors

$$\frac{q_i}{p_i} = \frac{\Phi(\sqrt{N_i} \theta_i^* / \sigma_i - T_i(y_{obs}))}{\Phi(-T_i(y_{obs}))}$$

are determined by the test statistics $T_i(y_{obs})$ and the noncentrality parameters $\lambda_i = \sqrt{N_i} \theta_i^* / \sigma_i$. When combined with the prior probabilities π_i that $\theta_i = \theta_i^*$, they determine the b-values needed to implement the rule r_B .

Table 1 illustrates the differences between the Bonferroni, frequentist FDR and Bayes FDR rejection rules when applied to hypothetical data for $m = 60$ hypotheses using the standardized mean statistics $T_i(Y)$, $i = 1, \dots, 60$. The table shows results for the eight hypotheses with p-values $p_i \leq \alpha = .05$. None of the hypotheses are rejected by the Bonferroni rule r_{BF} , which requires all p-values to satisfy $p_i \leq \alpha/60 = .0008$. In contrast, the frequentist FDR r_F rejects hypotheses H_1, \dots, H_4 , since the largest value of i for which $p_i \leq i\alpha/60$ is $i = 4$. Finally, the Bayes FDR rule r_B rejects only H_3 and H_4 , since rejection depends not only on the p-values but also on the study power and the prior probabilities that the hypotheses are false.

Table 1 here

Example 2: Hypothesis tests for regression coefficients in generalized linear models (GLMs). A more complex example occurs when the m hypotheses concern the coefficients in a regression model. To describe it, we assume that $Y = (Y_1, \dots, Y_N)$ is a vector of N independent random variables, each with a distribution in the exponential family specified by a GLM (McCullagh & Nelder, 1991). Specifically, the density function for Y_j takes the form

$$f_j(y; \theta) = \exp \left\{ (y\theta - b_j(\theta)) / \phi_j + c_j(y) \right\} \quad (10)$$

for some specific functions $b_j(\cdot)$ and $c_j(\cdot)$, with ϕ_j known, $j = 1, \dots, N$. The mean and variance of Y_j are $E[Y_j] = \mu_j = b'_j(\theta)$ and $\sigma_j^2 = b''_j(\theta)\phi_j$. The mean μ_j depends on the parameter θ through a function $\mu_j = g^{-1}(\eta_j)$, where $\eta_j = \sum_{i=0}^m x_{ji}\theta_i$ is a linear predictor involving a vector $x_j = (x_{j0}, \dots, x_{jm})$ of $m + 1$ known covariates, and $\theta = (\theta_0, \theta_1, \dots, \theta_m)$ lies in a compact open subset $\Omega \subset R^{m+1}$. We take $x_{j0} = 1$ to include an intercept θ_0 . Our objective is to test the m hypotheses $H_i : \theta_i = 0$ against specific alternatives $H_i^A : \theta_i = \theta_i^* \neq 0$, with $\theta_{i'}$, $i' \neq i$, unspecified, $i = 1, \dots, m$. We shall base our test statistic $T_i(Y)$ on the i^{Th} component $\hat{\theta}_i$ of the maximum likelihood estimate

$$\hat{\theta} = \sup_{\theta \in \Omega} L(\theta; y_{obs}) \quad \text{where } L(\theta; y) = \prod_{j=1}^N f_j(y; \theta),$$

and $f_j(y; \theta)$ is given by (10). In general, the exact distribution of $\widehat{\theta}_i$ for a given true value θ cannot be specified in closed form. Instead we shall approximate the probabilities p_i and q_i by invoking asymptotic maximum likelihood theory. Specifically, we assume that the marginal distribution of $\sqrt{N}\widehat{\theta}_i$ is approximately Gaussian with mean θ_i and variance τ_i^2 , where τ_i^2 is the i^{th} diagonal entry of the inverse of the observed information matrix $I(\theta) = -\partial^2 \log L(\theta; y_{obs}) / \partial \theta^2$. We take our test statistic as $T_i(Y) = \sqrt{N}\widehat{\theta}_i \tau_i^{-1}$, where τ_i is evaluated at $\theta = \widehat{\theta}$. The approximate two-sided p-value is $p_i = 2\Phi[-T_i(y_{obs})]$ and $q_i = \Phi\left[\sqrt{N}\theta_i^* / \tau_i - T_i(y_{obs})\right]$. The b-values can now be computed as in the previous example.

5. Application of Bayes rejection rule to data on breast cancer and genetic variation in repair of double-strand DNA breaks. Kuschel *et al.* (2002) used genotypes of 2200 female breast cancer cases and 1800 control women to evaluate $m = 19$ associations between breast cancer risk and single-nucleotide polymorphisms (SNPs) in seven genes involved in the repair of double-strand DNA breaks. Specifically, the authors tested the null hypothesis of no association with breast cancer for genotypes of 15 SNPs in the seven genes and for phased haplotypes containing SNPs within four of the genes (Table 2). They highlighted two SNPs in the gene XRCC3 (hypotheses H_{10} and H_{11}), one SNP in each of the genes XRCC2 and LIG4 (H_{13} and H_{15}), and a haplotype analysis involving $n = 8$ haplotypes of XRCC3 (H_{19}). Wacholder (2004) computed FPRPs for these five hypotheses. To do so, the authors determined q-values based on specified alternatives to each of the five hypotheses. They also used biological evidence and previous study findings to assign a range of prior probabilities that each of the hypotheses is false.

Table 2 here

With maximal FWER and FDRs set at $\alpha = .05$, only the hypothesis of no association between breast cancer and XRCC3 haplotypes (H_{19}) has a p-value $p \leq \alpha/19 = .0026$ needed to meet the Bonferroni criterion. Similarly, the frequentist FDR rule r_F rejects only H_{19} . Table 2 shows b-values

for all 19 hypotheses based on the alternative hypotheses specified by Wacholder *et al.* (2004) and three sets of prior probabilities. It can be seen that when the prior probability against each hypothesis is as high as 10%, hypotheses H_{16} and H_{19} are rejected. With a prior probability of .001, only hypothesis H_{19} meets the Bayes criteria.

6. **Discussion.** This paper defines a Bayes FDR for the multiple testing problem and relates it to the FPRPs described by Wacholder *et al.* (2004). The Bayes FDR is shown to be the mean FPRP among rejected hypotheses. Thus it can be controlled at level α by requiring all significant findings to have FPRPs no greater than α . Equivalently, significant findings must have posterior odds in their favor of at least $(1 - \alpha) / \alpha$. When $\alpha = .05$, for instance, significant findings must have posterior odds of at least 19:1.

There are some strengths and limitations to this Bayesian approach to multiple testing. In its favor, the approach allows a systematic accrual of evidence for rejecting a given null hypothesis. The posterior odds in favor of rejection, obtained after analyzing the data at hand, can form the prior odds for use in future studies. In this way, evolving data from ongoing studies can be combined in a transparent way to provide a summary measure of the total body of evidence for or against a finding. A limitation of the approach is the difficulty in choosing appropriate prior probabilities for the truth or falsity of a given hypothesis. When many null hypotheses are tested, the prior probability that any one is false is likely to be small. For large studies having close to 100% probability of rejecting the hypothesis when it is false, the magnitude of this prior probability plays a major role in interpreting the findings. Specifically, the posterior odds in favor of rejecting the hypothesis are approximately π/p , where π is the prior probability and p is the p-value. To achieve at least 2:1 posterior odds in favor of rejection, p must be less than $\pi/2$. Thus the choice of prior odds is key in determining the importance of a finding (see Matullo *et al.* (2005) for discussion of this issue in the context of gene-environment

interactions) . Since the choice of prior odds is somewhat subjective, interpretation of findings may be controversial. Still, the formalism involved in the Bayes FDR makes the needed judgement calls explicit.

In summary, the Bayesian approach to multiple testing outlined here provides an additional way to weigh the results of studies involving multiple testing, which can be used in combination with frequentist approaches such as the one proposed by Benjamini & Hochberg (1995).

References

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological*, 57(1), pp. 289-300.

Benjamini, Y. & Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), pp. 1165-1188.

Kuschel, B., Auranen, A., McBride, S., Novik, K. L., Antoniou, A., Lipscombe, J. M., Day, N. E., Easton, D. F., Ponder, B. A., Pharoah, P. D., & Dunning, A. (2002) Variants in DNA double-strand break repair genes and breast cancer susceptibility. *Human Molecular Genetics*, 11(12), pp. 1399-1407.

Matullo, G., Berwick, M., & Vineis, P. (2005) Gene-environment interactions: How many false positives? *Journal of the National Cancer Institute*, 97(8), pp. 550-551.

McCullagh, P. & Nelder, J. A. (1991) *Generalized Linear Models* (2nd Ed), (New York: Chapman & Hall).

Storey, J. D. & Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences U S A*, 100(16), pp. 9440-9445.

Thomas, D. C. & Clayton, D. G. Betting odds and genetic associations. (2004) *Journal of the National Cancer Institute*, 96(6), pp. 421-423.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman. N. (2004) Assessing

the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6), pp. 434-442.