

# Supplemental appendices

## A Testing the null hypothesis of ‘normal’ performance

For any individual patient, the observed outcome is binary: he or she either dies in hospital following surgery, or survives and is discharged. Consider a single provider and denote the response for the  $j^{\text{th}}$  patient undergoing surgery with that provider by  $Y_j$ , where  $j = 1, \dots, n$ . A death corresponds to  $Y_j = 1$  and a survival to  $Y_j = 0$ . Let  $H = 0$  if the null hypothesis that the provider’s performance is on target holds (in this case that its underlying RAMR is the same as the SWMR) and  $H = 1$  if the null hypothesis is false. Under the null

$$Y_j|H = 0 \sim \text{Bernoulli}(p_j); \quad j = 1, \dots, n \quad (1)$$

where  $p_j$  denotes the ‘normal’ probability of death for any patient with the same underlying level of risk (due to various factors over which the provider has no control). It is assumed that all  $p_j$ ’s are known, while in reality of course they have been estimated from a model.

The provider’s observed number of deaths is  $O = \sum_{j=1}^n Y_j$ . Similarly, its ‘expected’ death count is the sum of the expected probabilities over all of its patients:  $E = \sum_{j=1}^n p_j$ . If  $H = 0$ , the expectation of  $O$  is  $E$  while, if the patient outcomes are independent, its variance is

$$\text{Var}(O|H = 0) = \sum_{j=1}^n \text{Var}(Y_j) \quad (2)$$

$$= E - \sum_{j=1}^n p_j^2 \quad (3)$$

Since post-operative mortality rates are low overall, the factor  $\sum_{j=1}^n p_j^2$  is small and is ignored. The distribution of  $O$  under the null is then approximated by  $O|H = 0 \sim \text{Poisson}(E)$ . This approximation over-estimates the true variance by a very small amount.

The alternative hypothesis is that the provider’s RAMR is some multiple (which may be greater than or less than 1) of the SWMR. The general framework is therefore

$$O \sim \text{Poisson}(rE) \quad (4)$$

where  $r$  is the ‘relative risk’ associated with that provider. If  $H = 0$  then  $r = 1$ , while if  $H = 1$ ,  $r$  takes some other value.

Byar’s approximation provides a confidence interval for  $rE$ , the expectation of  $O$  [1]. Dividing this by the known  $E$  and multiplying by the SWMR, a confidence interval for the underlying RAMR is obtained, centered around the observed RAMR. The null hypothesis is rejected in favour of the alternative if this interval does not include

the SWMR (equivalently, if the confidence interval for  $r$  does not include 1). This is the approach used by the New York State Department of Health.

Our preferred formulation involves instead calculating how plausible each observed RAMR is, assuming that the target has been met. Consider the case where we wish to test if the provider's rate is significantly worse than expected. The hypothesis test is then one-sided, with the alternative being  $r > 1$ . A suitable  $p$ -value for this hypothesis test is the following *mid*  $p$ -value:

$$p(x) = P(O > x|H = 0) + 0.5P(O = x|H = 0) \quad (5)$$

where  $x$  is the observed value of the random variable,  $O$ .

We might also be interested in whether the provider's risk-adjusted rate is significantly *lower* than the target. A suitable  $p$ -value is  $\tilde{p}(x) = 1 - p(x)$ . Alternatively, the 2-sided  $p$ -value might be preferred, which is defined as  $2 \times \min(p(x), \tilde{p}(x))$ . Small values of this are associated with potential outliers that may have either higher or lower rates than expected.

Each test is carried out on each of  $m$  providers independently, resulting in a list of  $m$   $p$ -values.

## B Construction of funnel plot prediction lines

The lower prediction limits corresponding to a one-sided mid  $p$ -value threshold of  $\alpha$  are plotted as follows:

1. Find the largest integer,  $x^*$ , such that  $P(O < x) \leq \alpha$
2. Set  $\tilde{x} = x^* - 0.5 + c$ , where  $c = \frac{\alpha - P(O < x^*)}{P(O = x^*)}$
3. Plot the point  $(E, \frac{\tilde{x}}{E} \times SWMR)$
4. Repeat for many values of  $E$

The upper limits are obtained by replacing  $\alpha$  by  $1 - \alpha$  in the above algorithm.

## C Bayesian interpretation of the FDR

Here we outline the proof given by Storey [2] that the FDR for a fixed significance region can be written as the posterior probability that a null hypothesis is true given that it was rejected by the test.

Using basic laws of probability, the pFDR is seen to be

$$pFDR \equiv E(F/S|S > 0) \quad (6)$$

$$= \sum_{k=1}^m E(F/S|S = k)P(S = k|S > 0) \quad (7)$$

$$= \sum_{k=1}^m \frac{1}{k} E(F|S = k)P(S = k|S > 0) \quad (8)$$

For  $m$  independent and identical hypothesis tests,  $F|S = k$  has a binomial distribution with index  $k$ . Let  $H_i = 0$  if the  $i^{\text{th}}$  null hypothesis is true and denote the corresponding test statistic by  $T_i$  and the critical region by  $\Gamma$ . The probability parameter of the binomial distribution is intuitively  $P(H_i = 0|T_i \in \Gamma)$ , i.e. the probability that the null hypothesis is true given that it is rejected. This is the same for all  $i$ , and so the index can be dropped. The expectation of  $F|S = k$  is therefore  $k \times P(H = 0|T \in \Gamma)$ . Expression (8) then simplifies to

$$pFDR = \sum_{k=1}^m P(H = 0|T \in \Gamma)P(S = k|S > 0) \quad (9)$$

$$= P(H = 0|T \in \Gamma) \quad (10)$$

## D Derivation of Benjamini & Hochberg's algorithm

A formal proof that the algorithm controls the FDR at level  $\alpha^*$  can be found in [3]. Here we present only a heuristic justification.

For large  $m$ , the FDR is approximately equal to  $E(F)/E(S)$ . Say we reject all null hypotheses corresponding to  $p$ -values less than or equal to some threshold,  $t$ . Then  $E(S)$  is simply estimated by the observed number of rejected hypotheses,  $S(t)$ , while  $E(F)$  is equal to  $m_0 \times t$  (since each of the  $m_0$  null  $p$ -values follows a *Uniform*(0, 1) distribution and therefore has probability  $t$  of being  $\leq t$ ).

Defining  $\pi_0$  to be  $m_0/m$ , the proportion of hypotheses tested that are truly null, we then see that the FDR corresponding to this particular threshold is estimated by

$$FDR(t) = \frac{\pi_0 m t}{S(t)} \quad (11)$$

The aim is to determine the maximum value of  $t$  such that the FDR is less than or equal to  $\alpha^*$ , i.e. such that (rearranging (11))

$$t \leq \frac{S(t)}{m} \frac{\alpha^*}{\pi_0} \quad (12)$$

If  $t$  were set to the smallest  $p$ -value,  $p_{(1)}$ , then of course  $S(t)$  would be 1, while in general if  $t$  were set to  $p_{(i)}$  then  $S(t)$  would equal  $i$ . The desired critical FDR threshold is therefore the maximum  $p_{(i)}$  such that

$$p_{(i)} \leq (i/m)(\alpha^*/\pi_0) \quad (13)$$

In practice  $\pi_0$  is unknown and is difficult to estimate. It is therefore taken to be unity for the purposes of the algorithm. This is equivalent to constraining the FDR to be no greater than  $\alpha^*$  times the true value of  $\pi_0$ , and so is slightly conservative in the presence of genuine extremes.

Equivalently, the algorithm can be informally derived from a Bayesian perspective using Bayes' rule:

$$P(H = 0|T \in \Gamma) = \frac{P(T \in \Gamma|H = 0)P(H = 0)}{P(T \in \Gamma)} \quad (14)$$

The probability that  $H = 0$ , i.e. our prior probability that the null hypothesis holds for some provider, before observing the test statistic, is  $\pi_0$ .  $P(T \in \Gamma|H = 0)$  is  $t$  if the  $p$ -value threshold corresponding to a critical region of  $\Gamma$  is  $t$ , while  $P(T \in \Gamma)$  is estimated by the proportion of all hypotheses that are rejected,  $S(t)/m$ . Plugging these estimates into (14) we again arrive at (11).

## References

- [1] N E Breslow and N E Day. *Statistical methods in cancer research*, volume 2: The design and analysis of cohort studies. IARC, Lyon, 1987.
- [2] J D Storey. The positive false discovery rate: A Bayesian interpretation and the  $q$ -value. *Annals of Statistics*, 6:2013–2035, 2003.
- [3] Y Benjamini and Y Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Statist Soc B*, 57:289–300, 1995.