Geospatial Health 4(2), 2010, pp. 219-229

Multiple testing in disease mapping and descriptive epidemiology

Dolores Catelan[1,2], Annibale Biggeri[1,2]

[1]*Department of Statistics "G. Parenti", University of Florence, Florence, Italy;* [2]*Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Florence, Italy*

Abstract. The problem of multiple testing is rarely addressed in disease mapping or descriptive epidemiology. This issue is relevant when a large number of small areas or diseases are analysed. Control of the family wise error rate (FWER), for example via the Bonferroni correction, is avoided because it leads to loss of statistical power. To overcome such difficulties, control of the false discovery rate (FDR), the expected proportion of false rejections among all rejected hypotheses, was proposed in the context of clinical trials and genomic data analysis. FDR has a Bayesian interpretation and it is the basis of the so called q-value, the Bayesian counterpart of the p-value. In the present work, we address the multiplicity problem in disease mapping and show the performance of the FDR approach with two real examples and a small simulation study. The examples consider testing multiple diseases for a given area or multiple areas for a given disease. Using unadjusted p-values for multiple testing, an inappropriately large number of areas or diseases at altered risk are identified, whilst FDR procedures are appropriate and more powerful than the control of the FWER with the Bonferroni correction. We conclude that the FDR approach is adequate to screen for high/low risk areas or for disease excess/deficit and useful as a complementary procedure to point estimates and confidence intervals.

Keywords: descriptive epidemiology, disease mapping, small areas, multiple testing, false discovery rate.

Introduction

Atlases of mortality and morbidity, as well as a number of descriptive epidemiological studies, present long lists of relative risks and associated hypothesis tests that aim to detect deviations from local or national averages. Risk estimates and p-values may be reported for a given disease by a set of areas where the null hypothesis is no departure from the reference rate (e.g. regional or national). Alternatively, we may study the whole spectrum of diseases for a given area and be interested in assessing departure from a set of reference rates.

In such epidemiological investigations, estimation and prediction are the main issues. However, when considering public health implications of disease mapping, decision rules have been discussed (Richardson et al., 2004) but, in practice, only p-values are used to scrutinize lists of relative risks. This raises the question about how to communicate uncertainty and report scientific results. It is well known how easy p-values and confidence intervals (CI) can be misunderstood (Goodman, 2001; Sterne and Davey Smith, 2001). Notwithstanding, current reporting practice in descriptive epidemiology is to ignore multiple comparisons (see the influential paper of Rothman, 1990), while, in functional genomics data analysis and other "high-throughput" biological areas of applications, control of the false discovery rate (FDR) is popular because thousands of genes are simultaneously evaluated and there is no interest yet in estimation of quantitative effect measures (Storey, 2003; Wakefield, 2008).

Indeed, little work has been done on FDR in the field of geographical epidemiology. A PubMed search for "disease mapping, spatial mapping, spa-

Corresponding author:
Dolores Catelan
Department of Statistics "G. Parenti", University of Florence
Viale Morgagni, 59 - 50134 Florence, Italy
Tel. +39 055 4237472; Fax +39 055 4223560
E-mail: catelan@ds.unifi.it

tial epidemiology, FWER, FDR, multiple comparison" produced very few pertinent citations (Richardson et al., 2004; Best et al., 2005; Castro and Singer, 2006; Goovaerts et al., 2007; Catelan and Biggeri, 2008).

One explanation is that Bayesian approaches to smooth relative risk estimates may be interpreted as a solution to the problem (Clayton and Kaldor, 1987). Bayesian posterior estimators minimize a squared loss function over the whole set of areas and in this respect they are superior to maximum likelihood estimates (Efron and Morris, 1973). Bayesian estimates are also used in epidemiological applications where long lists of relative risks are compared (Greenland and Robbins, 1991; Carpenter et al., 1997). However the two approaches, FDR and Bayesian smoothing, do not necessarily produce similar inferences. Moreover, in studies such as disease mapping with a collection of parameters of interest, there are different inferential goals: estimating relative risks, their distribution, their ranks, and testing multiple hypotheses (Shen and Louis, 1998). There is no unique optimal solution for all those inferential tasks in the sense that there is no one estimator that is superior on all accounts. The approaches proposed in the literature have mainly been directed towards risk estimation. Heterogeneity variance and ranking have been addressed in only a few cases (Goldstein and Spiegelhalter, 1996; Pennello et al., 1999; Böhning et al., 2004; Catelan and Biggeri, 2008). With regard to hypothesis tests, posterior probabilities derived from hierarchical Bayesian models or classification probabilities from mixture models have been used (Biggeri et al., 2000; Militino et al., 2001; Best et al., 2005). However, all these approaches fail to address the multiple comparison problems because an appropriate Bayesian approach would consist of a tri-level hierarchical Bayesian model for the estimation of the posterior probability of a given disease/area belonging to the set of null or alternative hypotheses (Müller et al., 2006; Catelan et al., 2009).

Notwithstanding this, there is a field where multiple comparison is considered: cluster detection in spatial and spatio-temporal surveillance. There control of the family wise error rate (FWER) is applied (Kulldorff, 2001; Frisén, 2003). Bonferroni correction is the most popular FWER control: for fixed level α of probability of type I error, it consists in doing m hypothesis tests controlling the significance level per test at $\alpha^*=\alpha/m$. This approach is extremely conservative whenever the number of tests is large. Rolka et al. (2007) discussed the cost in sensitivity of adopting a FWER approach for monitoring and mentioned the FDR control in epidemiological surveillance. FDR is the expected value of the rate of false positives among all rejected hypotheses and was introduced, in the context of clinical trials by Benjamini and Hochberg (1995) to improve the statistical power of the test. FDR has a Bayesian interpretation and it is the basis of the so called q-value (Storey, 2003), the Bayesian counterpart of the p-value. The FWER and the FDR are based on different philosophies. The former is appropriate when we are performing multiple tests of the same null hypothesis. On the other hand, if we are testing different null hypotheses the latter is more appropriate and allows for an easier communication of the related uncertainty. In the last years, FDR control has been proposed for the screening of "hot spots" (Castro and Singer, 2006) and for profiling health care providers (Ohlssen et al., 2007; Jones et al., 2008).

In the present paper, we address and justify the FDR approach in the context of descriptive epidemiology with two real examples and via a small simulation study.

Materials and methods

Motivating examples

The first example is based on the report "Environment and Health in Sardinia" (Biggeri et al., 2006) commissioned by the Secretary of Health, Hygiene and Social Welfare of the Region to screen the health status of the resident populations in 18 *a priori* identified areas at high environmental pressure by industrial mining and military activities. For each area, 36 ICD9 disease codes with respect to mortality and 48 disease codes with respect to hospital admission were analysed resulting in 1,512 estimates of relative risk for each sex. We are in a "large table" context as defined in Carpenter et al. (1997) and Law et al. (2001). For the purpose of this work, we extracted 29 age- and deprivation-adjusted (Carstairs, 1995) standardised mortality ratios (SMR) for the period 1997-2001 that referred to mutually exclusive ICD9 disease codes relative to male residents in the industrial area of Portoscuso, one of the 18 analysed areas. The goal was to identify the diseases for which the population demonstrated a risk that was divergent from the regional mean.

The second example refers to geographical descriptive epidemiology and uses data from the "Atlas of Mortality in Tuscany" (University of Pisa, 2001). Lung cancer death certificates were considered for male residents in the 287 municipalities of the Tuscany region (Italy) for the period 1995-1999. Following indirect standardisation and classifying of the population by 18 classes of age (0-4 years, 5-9 years etc. up and including "85 years or more"), a set of reference rates (Tuscany, 1971-1999) were used to compute the expected number of cases for each municipality. The goal was to identify municipalities with divergent risk from the regional mean.

While controlling for the uncertainty related to multiple testing, the two examples belong to different camps of epidemiological surveillance: identifying the diseases at altered risk in a given area vs. identifying the areas at altered risk for a given disease. In both situations, we wanted to test different null hypotheses and our decision, based on the whole set of estimates of relative risks, was not erroneous even if some of the null hypotheses were falsely rejected. Indeed, in the first example, to define an area as being at altered risk on the basis of, for example, 10 diseases in excess/deficiency are not erroneous even if some of the rejected null hypotheses are erroneously rejected. In the second example related to disease mapping, the erroneous rejection of the null hypothesis for some municipalities does not challenge the results of the whole descriptive analysis whose aim is to assess heterogeneity of risk in the entire study region.

In both cases, it is more appropriate to control the expected number of false positives over the total number of rejected hypotheses (FDR), than strictly control the probability of at least one false positive among all tests (FWER) (Benjamini and Hochberg, 1995).

Methods

Table 1 reports the possible outcomes from the *m* tests where *m* is known *a priori* and R, the number of rejected hypotheses, is the only observable random variable. In the context of multiple inferences, we traditionally control the error rate at a desired level while maintaining the power of each test as much as possible. The commonly controlled quantity is FWER, the probability of which yields one or more false positives out of all tests. The most commonly used method is the Bonferroni correction, i.e. if we set the probability of type I error at α and m tests are performed, each test is controlled at the level $\alpha^* = \alpha/m$ (Bonferroni, 1936). This guarantees that the probability of a false positive is at maximum equal to α.

Table 1. Possible outcomes from hypothesis tests.

	Accept H_0	Reject H_0	
True H_0	U	V	M_0
Not true H_0	T	S	$m-M_0$
	m-R	R	m

Formally, in terms of Table 1, we can define the FWER = Pr (V ≥1). However, there are situations (see the motivating examples) in which this quantity is not of real interest and the attention moves to the expected proportion of error among all rejected hypotheses. Benjamini and Hochberg (1995) called this quantity the false discovery rate (FDR). Following the notation of Table 1, the FDR can be formally defined as:

$$\mathrm{FDR} = \mathrm{E}\left[\frac{V}{R \cup 1}\right] = \mathrm{E}\left[\frac{V}{R} \middle| R > 0\right] \Pr(R > 0),$$

where "E[.]" stands for expected value and "R∪1" is the maximum between "R" and 1 which guarantees that FDR = 0 when no hypothesis is rejected. The FDR is thus defined as the expected value of the ratio between the number of null hypotheses erroneously rejected and the total number of rejected hypotheses. If $M_0 = m$, the null hypothesis is true for all tests, then FDR = FWER, because

$$E\left[\frac{V}{R \cup 1}\right] = \Pr\ (R > 0).$$

Otherwise it is lower. The larger the number of true rejections is, the larger the gain in power is.

Storey (2003) defined the positive FDR (pFDR) as:

$$\text{pFDR} = \text{E}\left[\frac{\text{V}}{\text{R}} \,\middle|\, \text{R} > 0\right],$$

where "positive" denotes that we are conditioning the occurrence of at least one rejection. We cannot control positive FDR because if $M_0 = m$ then pFDR = 1. This is natural since any rejection must be false if all null hypotheses are true. However for a given significance level, i.e. a given rejection region, a conservative null estimate of pFDR can be obtained.

The pFDR can be derived as a posterior Bayesian probability, i.e. the posterior probability of the null given that the test statistics fall in the rejection region. This derivation assumes that the tests are exchangeable in the sense that we have no reason to believe that the probability of the null should be greater for some tests (in detail, let $H_i = 0$ denote the event that the null is true for the *i*-th test, then assume the H_i are independent and identically distributed (*iid*) Bernoulli random variables and the test statistics are *iid* $T_i|H_i \sim (1\text{-}H_i)\text{x}F_0 + H_i\text{x}F_1$ for some null and alternative distribution *F*) (Storey, 2003).

The pFDR has a connection with classification theory and can be used to define the q-value, the Bayesian counterpart of the p-value. The pFDR approach uses all information from prior probability of the null (Storey, 2002) and from the data and is thus more powerful than the FDR (see Storey, 2003, for a complete description of the properties of pFDR).

If we define for a given type I error probability α, a test statistic T and a rejection region G, the pFDR can be written as:

$$\text{pFDR}(\Gamma) = \Pr(H_0 \mid T \in \Gamma) =$$

$$\frac{\pi_0 \Pr(\text{T} \in \Gamma \mid H_0)}{\pi_0 \Pr(T \in \Gamma \mid H_0) + \pi_1 \Pr(T \in \Gamma \mid H_1)}$$

that is a posterior Bayesian probability. Note that the probability to observe a certain value for the test statistic T is equal to a mixture of the distribution under null (H_0) and alternative (H_1) hypothesis:

$$\Pr(T \in \Gamma) = \pi_0 \Pr(T \in \Gamma \mid H_0) + \pi_1 \Pr(T \in \Gamma \mid H_1),$$

where π_0 is the *a priori* probability that the null hypothesis is true and $\pi_1 = 1\text{-}\ \pi_0$.

If we consider the probability of values equal or more extreme than T_{obs} (observed value of the test statistic) we have:

$$\text{pFDR}\{(T \geq T_{obs})\} =$$

$$\frac{\pi_0 \Pr(T \geq T_{obs} \mid H_0)}{\pi_0 \Pr(T \geq T_{obs} \mid H_0) + \pi_1 \Pr(T \geq T_{obs} \mid H_1)}$$

$$= \Pr\ (H_0 \mid T \geq T_{obs}).$$

This posterior probability is defined as the *q-value* and represents the Bayesian analogue of the p-value – $\Pr\ (T \geq T_{obs}|H_0)$. In other words, the p-value is a measure of the strength of the empirical evidence against the null hypothesis (it is the minimum type I error that we can incur when we reject the null hypothesis on the basis of the observed value of the test statistics T). The q-value is the minimum pFDR that we can incur when we reject the null hypothesis on the basis of the observed or more extreme values of the test statistics T. The q-value is a measure that takes into account for the multiplicity of the tested hypotheses and represents how far we are from the null hypothesis on the basis of observed data. Being a posterior probability, the q-value takes into account not only the null, but also the alternative hypothesis

(Goodman, 2001). Algorithms for the evaluation of pFDR and q-value are discussed in Storey (2002). Additional details are provided in the Appendix.

Results

Example 1

From the report on high-risk areas of the Sardinian region, we extracted 29 age- and deprivation-adjusted SMRs for the period 1997-2001 for male residents in the industrial area of Portoscuso (Biggeri et al., 2006). In Table 2, we report the observed and expected number of cases, including p-values and q-values for the 10 diseases with associated p ≤0.10. Setting α at 10%, Bonferroni's correction gives a significance level α^* = 0.0035 (0.10/29 the number of causes of death analyzed): only three p-values are below α^*. By thresholding q-values at 10%, we identified five diseases. Of note, whilst the p-value is the minimum rate of false positive, in which we incur rejecting the null hypothesis, the q-value represents the minimum rate of false "discoveries", which is the proportion of erroneously rejected tests over all the rejected tests. It means that over the five diseases declared different from the null by the q-value, "on average" 0.10 x 5 is a false positive, even if we cannot identify which of them it is.

Table 2. Example 1 results. Observed and expected number of cases, p-value and q-value. Industrial area of Portoscuso, male death certificates 1997-2001. Reports on high risk areas in the Sardinian region, Italy. Diseases with FDR ≤10% are reported in bold.

Disease	Cases	Expected	p-value	q-value
Pneumoconiosis	112	30.47	**<0.0001**	**<0.0001**
Diabetes	15	33.09	**0.0005**	**0.0066**
Prostate cancer	15	32.63	**0.0007**	**0.0066**
Coronary heart disease	126	155.36	**0.0161**	**0.0979**
Lung cancer	136	109.63	**0.0169**	**0.0979**
Respiratory acute	48	33.85	0.0250	0.1095
Liver cirrhosis	30	44.03	0.0292	0.1095
Ill defined conditions	10	18.99	0.0302	0.1095
Respiratory chronic	45	60.45	0.0403	0.1299
Melanoma	0	3.00	0.0536	0.1554

Example 2

Considering the second example, i.e. lung cancer mortality in Tuscany at the municipality level, the SMRs exhibit a strong geographical heterogeneity (Fig. 1) and a wide range of values (Fig. 2) suggesting the presence of areas at high/low risk for the investigated disease. In Table 3, we report the observed and the expected numbers of cases, including p-values and q-values for the 11 municipalities with the lowest q-values. Fixing α at 0.10 Bonferroni's correction results in a threshold level α^* of 0.0004 (0.10/287, which is the number of municipalities) only two p-values are below α^*. Thresholding q-values at 10% identify six municipalities. Selecting p-values at α = 0.10 level of significance leads to a high number of rejections and, on the other side, the FWER control results in only two comparisons satisfying the significance level.

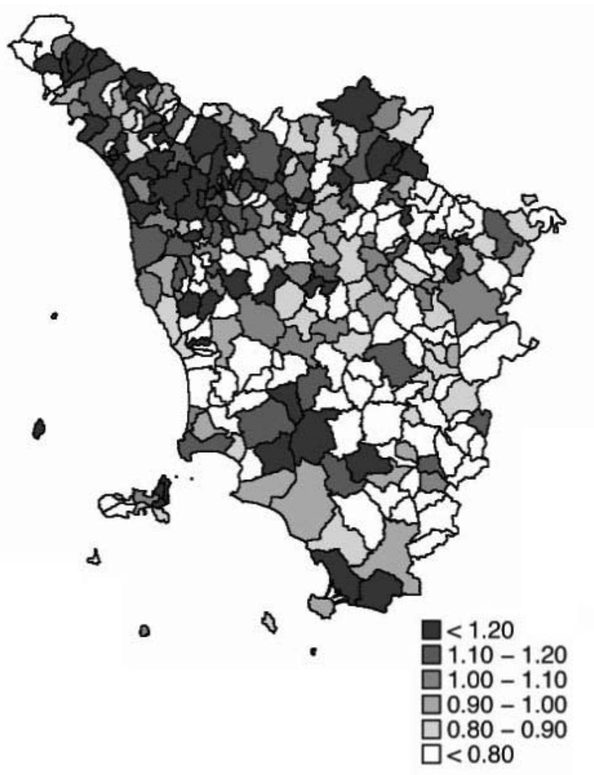

Fig. 1. Geographical distribution of relative risk (SMR) for male lung cancer mortality in Tuscany region, Italy, 1995-1999.

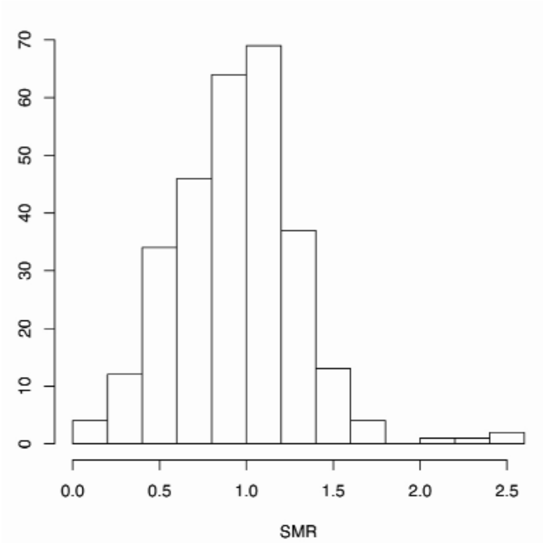

Fig. 2. Histogram of relative risks (SMR) for male lung cancer mortality in Tuscany region, Italy, 1995-1999.

Table 3. Example 2 results. Observed and expected number of cases, p-value and q-value. Male lung cancer mortality, Tuscany region, Italy, 1995-1999. Municipalities with FDR ≤10% are reported in bold.

ISTAT Municipality code	Cases	Expected	p-value	q-value
46033	205	138.80	<0.0001	0.0001
46005	108	73.53	0.0002	0.0248
45003	203	157.67	0.0006	0.0447
51020	3	13.87	0.0006	0.0447
46017	266	215.41	0.0009	0.0523
52035	8	20.95	0.0017	0.0766
47012	71	48.89	0.0032	0.1226
51001	6	16.91	0.0035	0.1226
48001	47	69.94	0.0041	0.1274
46002	34	19.60	0.0047	0.1293
52007	2	9.28	0.0065	0.1637

The estimates of the proportion p_0 of true null hypothesis proposed by Storey (2002) were close to one in the first example and 97% in the second example. They represent two different situations: in the first example we have a low-moderate number of hypotheses tests (below 50) and in the second we have a moderate-high number of hypotheses tests (between 200 and 500). In the last situation, we have enough data to consistently estimate p_0 and when the number of comparisons is small it has been suggested not to estimate p_0 and fix it to one (Ohlssen et al., 2007).

Split sample study and a small simulation study

To demonstrate the point that p-values actually lead to false positives regarding the hypotheses tested, we have used data from the second example. We took male lung cancer deaths over the 10-year calendar period 1996-2005 in the 287 Tuscan municipalities and analysed the whole 10-year period and two 5-year periods obtained summing up even or odd years. Of 68 (31+2, 30+5) discordant municipalities, i.e. with p-values below the two-sided 10% significance level in one split-sample and above the significance level in the other, 61 (90%) were not detected by thresholding q-values at 10%. Using p-values for the whole 10-year period, we would have declared up to 39 (18+21) over 68 (57%) municipalities as divergent with discordant p-values in the subsets (Table 4). The q-value approach leads to appropriately discarding the less reproducible results, which is not the case when p-values are used.

To check the assumption of uniformity of p-values under the null hypothesis, we simulated 1,000 datasets from the data of the second example, i.e. lung cancer among males in 287 municipalities of the Tuscany region. Using internal indirect standardization, values obtained over the 10-year period, 1996-2005, were assumed to be expected values. The marginal distribution of p-values over all the municipalities was roughly uniform. The percentage of p-values, less than or equal to 0.05 was 4.43%, i.e. slightly below the nominal value. The smaller the Poisson mean the more conservative the test statistic. Indeed in this data only 6.97% of municipalities had low expected counts under indirect standardisation, i.e. small population at risk. Our approach based on q-values should be considered with caution when the number of observations with expected counts less than five is greater than 20%.

Table 4. Male lung cancer mortality in Tuscany region, Italy, 1996-2005. Analysis of the whole data period and on split data by odd or even calendar year. Areas are cross-classified by p-values ≤0.10 and q-values.

	Even years (1996, 1998, 2000, 2002, 2004)			
	$p > \alpha$		$P \leq \alpha$	
	Odd years (1997, 1999, 2001, 2003, 2005)			
1996-2005	$p > \alpha$	$p \leq \alpha$	$p > \alpha$	$p \leq \alpha$
q >0.10	203	31	30	3
q ≤0.10	0	2	5	13
$p > \alpha$	191	15	14	1
$p \leq \alpha$	12	18	21	15

Discussion

Computation of the q-value is easy and essentially based on observed p-values. The procedure is robust and independent from the model used to obtain the p-values (Storey, 2002). The FDR approach has a strong advantage in terms of statistical power with regard to standard procedures for FWER control such as that of Bonferroni. Here we are concerned with controlling the rate of false discoveries, while at the same time maintaining acceptable power. If the study aims to explore any possible deviations from the null hypothesis, the reader should be advised that the proposed approach could still be weak in terms of the rate of false negatives. Eventually, controlling FDR at 20% may be appropriate.

We used a thresholding level of 10% for q-values. Recently Jones et al. (2008) discussed this approach in the context of health care provider profiling and noted that q-values can be screened in an exploratory manner without having a fixed significance level in advance. This is an advantage regarding the FDR controlling procedure, where we need to declare a cut-off point (Benjamini and Hochberg, 1995).

With regard to the estimate of π_0, i.e. the true null hypothesis proportion, we suggest exploring the stability of the estimate assuming different *a priori* beliefs (Storey, 2002) and to compare the results with the Benjamini-Hochberg approach that implicitly assumes the conservative prior of 1 for π_0 (see for example Jones et al., 2008). Benjamini et al. (2006) discussed several procedures that also are suitable when the number of tests is not large. Applying their approaches in the present settings produced results identical to those obtained by thresholding q-values.

It is not always easy to figure out p-values that have coverage probabilities that correspond to the nominal ones. With count data (e.g. number of deaths or cases of disease), a maximum likelihood estimate of relative risk is given by the ratio of the observed and expected number of cases using indirect standardisation. Exact p-values can be obtained directly from the Poisson distribution. It must be noted that in this case, one and two-sided p-values may coincide. Eventually, direct exploration of the likelihood profile for the rate ratio under the null hypothesis is necessary. Some would argue in favor of using mid-p adjustment (Jones et al., 2008). However the approach described applies to whatever model chosen to calculate p-values.

It can be argued that spatial dependence could be a problem with the FDR approach. We also applied the Yekutieli-Benjamini procedure, which is robust for violation of independence assumption (Yekutieli-Benjamini, 1999). As expected, we obtained a smaller number of discoveries (three in the first example and one area in the second example, thersholding at 0.10). That procedure, which is robust under dependence, behaved conservatively in our examples. We prefer to accept the independence assumption. The reader should note that it is reasonable to assume exchangeability in the context of disease mapping as demonstrated by Lawson et al. (2000). In that paper, a large simulation study showed favourable behaviour of spatially independent models.

Indeed, here information from the whole distribution of the test statistics, or from a set of them, was not considered. For example, in the disease mapping context, it is popular to make inference with regard to the relative risk of one area using the information coming from adjacent areas. Storey (2007) proposed an optimal discovery procedure

(ODP) to make multiple inferences when using non-independent observation. The ODP uses information coming from all the hypotheses tests and the significance of each test is given by the weighted average of p-values of all the tested hypotheses (a kind of "shrinkage" estimator). The ODP, introduced by Storey (2007) in the analysis of DNA microarray data, shares the same ideas, which underlie Bayesian methods, developed in the context of disease mapping (Clayton and Caldor, 1987; Besag et al., 1991).

In the present paper, we conservatively maintain the exchangeability assumption. A full Bayesian alternative is to develop a tri-level hierarchical Bayesian model to estimate the posterior probability of a given disease/area to belong to the set of the null hypothesis or alternative sets (Müller et al., 2006). Further developments will consider spatially structured priors in hierarchical Bayesian models for disease mapping (Catelan et al., 2009).

Other approaches have been proposed in the scientific literature for analysing "large tables". For example, Quantile-Quantile plots of empirical Bayes estimates with appropriate "guide rails" have been proposed for the screening of long lists of relative risks (Carpenter et al., 1997; Law et al., 2001). This graphical approach is appealing, but it is more appropriate in detecting the presence of a small number of outlying risks and may fail to detect extremes when a high number of altered diseases (within the area) or areas (for a given disease) is expected. Others have taken advantage of the distance among cases and have developed cluster detection methods in the context of epidemiological surveillance. There are applications using those methods on areal data like those used in disease mapping. Multiple testing is controlled by the FWER procedure (Kulldorff, 2001). Rolka et al. (2007) discussed the use of FDR in surveillance. However, the reader should be aware that in the present work, example 2 is in the context of a disease atlas where a cluster detection analysis would be inappropriate, because the size of the areas is too large.

Finally, when we are concerned with estimating the set of altered diseases/areas among all available data, the approach outlined in this paper is appropriate. If we were to be concerned only with the estimation of the relative risk of one disease/area there would be only one test statistic and no multiple testing problem would arise. In this case, assuming a Bayesian perspective, under a different prior in favour of the null hypothesis, posterior probabilities having the same interpretation of q-value can be derived as outlined by Goodman (2001).

Conclusions

Geographical epidemiology is essentially based on maps of relative risk or identification of clusters of disease cases. Specific investigations usually follow positive results. In disease surveillance, control of type I error is commonly pursued following FWER procedures (a discussion on alternatives to improve the sensitivity of these techniques can be found in Rolka et al. (2007). In current practice, there is no formal adjustment in disease mapping.

We present two examples in which inconsistencies due to multiple inferences are present, i.e. to evaluate different diseases in a given area or the same disease across different areas. We considered the Bayesian alternative to p-value, i.e. the q-value (Storey, 2003), to identify extreme diseases or areas. Bayesian posterior probabilities are more appropriate as illustrated by the examples and may be more useful to communicate results to lay people.

We suggest the use of FDR control procedures to screen for high risk areas or for excess diseases. This approach is appropriate for coping with multiple testing and overcomes the low power of the traditional FWER control procedures. FDR control is suggested as complementary to the use of point estimate and confidence intervals.

Appendix

The p-values are distributed as Uniform (0,1), under the null. The distribution of the ordered p-values is Beta (i,m+1-i) with expectation $i/(m+1)$. The smallest p-value over m tests is therefore distributed as follows:

$$F(p_{(1)}) \propto 1\text{-}(1\text{-}p_{(1)})^m \approx m \times p_{(1)}$$

when p is small, and the generic i-th ordered p-value as;

$$F(p_{(i)}) \propto 1\text{-}(1\text{-}p_{(i)})^m \approx \frac{m}{i} \times p_{(i)}\,.$$

Therefore, FDR can be controlled at level α sorting the p-values and taking as the cut-off critical value the one that corresponds to the largest p-value which satisfies:

$$p_{(i)} \leq \frac{i}{m}\ \alpha.$$

If i = 1, the smallest p-value, the procedure is equivalent to Bonferroni's correction.

Instead of looking for the cut-off corresponding to a desired level of FDR, we can explore all the test statistics, taking the observed test statistics as critical point. This is the way we explore the whole set of p-values and here we can use pFDR to derive a more appropriate quantity, the q-value.

The pFDR can be written in terms of a posterior probability as:

$$\text{pFDR} = \frac{\pi_0 \Pr(T \in \Gamma \mid H_0)}{\Pr(T \in \Gamma)} = \frac{\pi_0\ \text{p-value}}{\Pr(\text{rejection})}$$

that can be estimated, for a given rejection region, as

$$\text{p}\hat{\text{FDR}} = \frac{\pi_0\, p_{(i)}}{i/m}\ .$$

The q-value is therefore given by

$$\text{q-value} = \min\left[\frac{\pi_0\, p_{(i)}}{i/m};\ q\text{-value}_{(i+1)}\right].$$

The pFDR procedure depends on π_0. The Benjamini-Hochberg approach implicitly assumes $\pi_0 = 1$. This is conservative. Storey (2002) proposed an algorithm to estimate $\hat{\pi}_0$:

$$\hat{\pi}_0 = \frac{\#\ \text{p-values} > \lambda}{(1 - \lambda)\ m}$$

for fixed λ. A bootstrap method to select optimal λ is given by Storey (2002).

q-value is a posterior probability and can be expressed as

posterior odds = prior odds x Bayes factor.

Goodman (2001) provides examples of the Bayes factor which are functions of p-value. They are interpretable as a global minimum over the null hypothesis:

$$\text{minimum BF} = \exp\{-\Phi^{-1}\ (\text{p-value})^2/2\}$$

where Φ^{-1} is the inverse Normal and p-value is two sided.

For example, if we assume a prior odds for the null 9:1, then, to achieve at least a posterior odds of 1:4 (i.e. controlling FDR at 20%), we should have a minimum BF=1/36 (indeed, 1/4=9/1 x BF). The associated p-value is at most 0.0037, which is strong evidence against the null while controlling for multiple comparisons (Goodman, 2001).

Acknowledgements

We acknowledge: Project AGIRE POR Sardinia-Tuscany, European Union, Ministry of Economic Development, Ministry of Labour, Health and Social Policies, 2007-2008 (DC, AB); Ministry of Education, University and Scientific Research PRIN 2006131039 “Statistical Methods for Environmental Health Impact Assessment” 2006-2008 (DC); Ministry of Education, University and Scientific Research PRIN 2005134079 “Statistical Tools in System Biology” 2005-2007 (AB).

References

Benjamini Y, Hochberg Y, 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57, 289-300.

Benjamini Y, Krieger AM, Yekutieli D, 2006. Adaptive linear step-up false discovery rate controlling procedures. Biometrika 93, 491-507.

Besag J, York JC, Molliè A, 1991. A Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann Inst Statist Math 43, 1-59.

Best N, Richardson S, Thomson A, 2005. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 14, 35-59.

Biggeri A, Lagazio C, Catelan D, Pirastu R, Casson F, Terracini B, 2006. Environment and health in Sardinia. Epidemiologia e Prevenzione 30S, 1-96.

Biggeri A, Marchi M, Lagazio C, Böhning D, Martuzzi M, 2000. Non parametric maximum likelihood estimators for disease mapping. Stat Med 19, 2539-2554.

Bonferroni C, 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3–62.

Böhning D, Sarol J, Rattanasiri S, Viwatwongkasem C, Biggeri A, 2004. A comparison of non-iterative and iterative estimators of heterogeneity variance for the standardized mortality ratio. Biostatistics 5, 61-74.

Carpenter LM, Maconochie NES, Roman E, Cox DR, 1997. Examining associations between occupation and health by using routinely collected data. J Roy Stat Soc A 160, 507-521.

Carstairs V, 1995. Deprivation indices: their interpretation and use in relation to health. J Epidemiol Commun Health 49S, 3-8.

Castro MC, Singer BH, 2006. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. Geogr Anal 38, 180-208.

Catelan D, Biggeri A, 2008. A statistical approach to rank multiple priorities in environmental epidemiology: an example from high-risk areas in Sardinia, Italy. Geospat Health 3, 81-89.

Catelan D, Lagazio C, Biggeri A, 2009. Statistical approaches to environmental spatial surveillance. In: Proceedings of the Multiple Procedures Conference, 6-9 March 2009, Tokyo, Japan.

Clayton DG, Kaldor J, 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671-681.

Efron B, Morris C, 1973. Stein's estimation rule and its competitors: an empirical Bayes approach. J Am Stat Assoc 68, 117-130.

Frisén M, 2003. Statistical surveillance. Optimality and methods. Int Stat Rev 71, 403-434.

Goldstein H, Spiegelhalter DJ, 1996. League tables and their limitations: statistical issues in comparisons of institutional performance. J Roy Stat Soc A 159, 385-443.

Goodman SN, 2001. Of p-values and Bayes: a modest proposal. Epidemiology 12, 295-297.

Goovaerts P, Meliker JR, Jacquez GM, 2007. A comparative analysis of spatial statistics for detecting racial disparities in cancer mortality rates. Int J Health Geogr 6, 32-44.

Greenland S, Robbins J, 1991. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 2, 244-251.

Militino AF, Ugarte MD, Dean CB, 2001. The use of mixture models for identifying high risks in disease mapping. Stat Med 20, 2035-2049.

Müller P, Parmigiani G, Rice K, 2006. FDR and Bayesian multiple comparison rules. In: Proceedings Valencia/ISBA, 8th World Meeting on Bayesian Statistics. Benidorm, Alicante, Spain.

Jones HE, Ohlssen DI, Spiegelhalter DJ, 2008. Use of the false discovery rate when comparing multiple health care providers. J Clin Epidemiol 61, 232-240.

Kulldorff M, 2001. Prospective time-periodic geographical disease surveillance using a Scan Statistic. J Roy Stat Soc A 164, 61-72.

Law G, Cox DR, Machonochie N, Simpson J, Roman E, Carpenter L, 2001. Large tables. Biostatistics 2, 163-171.

Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, Schlattmann P, Divino F, 2000. Disease mapping models: an empirical evaluation. Stat Med 19, 2217-2241.

Ohlssen DI, Sharples LD, Spiegelhalter DJ, 2007. A hierarchical modelling framework for identifying unusual performance in health care providers. J Roy Stat Soc A 170, 865-890.

Pennello GA, Devesa SS, Gail MH, 1999. Using a mixed effects model to estimate geographic variation in cancer rates. Biometrics 55, 774-781.

Richardson S, Thomson A, Best N, Elliot P, 2004. Interpreting posterior relative risk estimates in disease mapping studies. Environ Health Persp 112, 1016-1025.

Rolka H, Burkom H, Cooper G, Kulldorff M, Madigan D, Wong WK, 2007. Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research need. Stat Med 26, 1834-1856.

Rothman K, 1990. No adjustments are needed for multiple comparisons. Epidemiology 1, 43-46.

Shen W, Louis TA, 1998. Triple-goal estimates in two-stage hierarchical models. J Roy Stat Soc B 60, 455-471.

Stern JAC, Smith DG, 2001. Sifting the evidence. What's wrong with significance tests? BMJ 322, 226-231.

Storey JD, 2002. A direct approach to false discovery rates. J Roy Stat Soc B 64, 479-498.

Storey JD, 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31, 2013-2035.

Storey J, 2007. The optimal discovery procedure: a new approach to simultaneous significance testing. J Roy Stat Soc B 69, 347-368.

Wakefield J, 2008. Reporting and interpretation in genome-wide association studies. Int J Epidemiol 37, 641-653.

Yekutieli D, Benjamini Y, 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plan Infer 82, 171-196.