BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Optimal likelihood-ratio multiple testing with application to Alzheimer's disease and questionable dementia

Donghwan Lee[1†], Hyejin Kang[2,3†], Eunkyung Kim[2,4], Hyekyoung Lee[2], Heejung Kim[2,4], Yu Kyeong Kim[2,5], Youngjo Lee[3,6*] and Dong Soo Lee[2,4,7*]

Abstract

Background: Controlling the false discovery rate is important when testing multiple hypotheses. To enhance the detection capability of a false discovery rate control test, we applied the likelihood ratio-based multiple testing method in neuroimage data and compared the performance with the existing methods.

Methods: We analysed the performance of the likelihood ratio-based false discovery rate method using simulation data generated under independent assumption, and positron emission tomography data of Alzheimer's disease and questionable dementia. We investigated how well the method detects extensive hypometabolic regions and compared the results to those of the conventional Benjamini Hochberg-false discovery rate method.

Results: Our findings show that the likelihood ratio-based false discovery rate method can control the false discovery rate, giving the smallest false non-discovery rate (for a one-sided test) or the smallest expected number of false assignments (for a two-sided test). Even though we assumed independence among voxels, the likelihood ratio-based false discovery rate method detected more extensive hypometabolic regions in 22 patients with Alzheimer's disease, as compared to the 44 normal controls, than did the Benjamini Hochberg-false discovery rate method. The contingency and distribution patterns were consistent with those of previous studies. In 24 questionable dementia patients, the proposed likelihood ratio-based false discovery rate method was able to detect hypometabolism in the medial temporal region.

Conclusions: This study showed that the proposed likelihood ratio-based false discovery rate method efficiently identifies extensive hypometabolic regions owing to its increased detection capability and ability to control the false discovery rate.

Background

Several multiple hypothesis testing methods have been proposed for use in neuroimaging studies. Bonferroni correction is the simplest but the most conservative method for controlling the family-wise error rate (FWER). However, it often fails to detect voxels with real activation or difference. As an alternative approach, the Benjamini and Hochberg [1] method for controlling the false discovery rate (FDR) was applied to neuroimaging studies by Genovese, Lazar and Nichols [2]. The FDR control gives statistically less conservative procedures than FWER. However, Cohen and Sackrowitz [3] proved that the Benjamini and Hochberg procedure is inadmissible under any loss function that is a linear combination of false discoveries and false non-discoveries. Given a fixed FDR, it is desirable to maximize the power by minimizing the false non-discovery rate (FNDR).

Recently, Lee and Bjørnstad [4] proposed a new multiple hypothesis test based on the likelihood-ratio-based FDR (LR-FDR). They showed that the problem of large-scale multiple testing is naturally expressed as an inference problem for finding the true discoveries. And they represented the underlying effects of interest by the (unknown)

* Correspondence: youngjo@snu.ac.kr; dsl@snu.ac.kr
†Equal contributors
[3]Data Science for Knowledge Creation Research Center, Seoul National University, Seoul, Korea
[2]Department of Nuclear Medicine, Seoul National University College of Medicine, Seoul, Korea
Full list of author information is available at the end of the article

discrete random variables. Statistical inferences are for two types of unknowns, namely parameters (fixed unknowns) and unobservables (random unknowns). Bjørnstad [5] showed that all information on parameter and unobservable data was in the extended likelihood, such as the h-likelihood [6]. Lee, Nelder and Pawitan [7] extensively introduced a random effect analysis using the extended likelihood. More recently, Lee and Bjørnstad [4] showed how the extended likelihood can be used to derive their proposed LR-FDR method. This method is optimal when (a) determining the order in which the test results can be called significant and (b) controlling error rates given this order. Provided an assumed statistical model is true, the likelihood exploits all information in the data to provide the most efficient testing. Therefore, it is important to search for the best-fitting model enhancing the performance of a multiple hypothesis test. The likelihood approach provides various well-developed model-checking and model-selection procedures.

In reviewing existing multiple tests, Efron [8] began by summarizing statistics such as *p*-values [1] and test statistics [8]. He then described how to find a single-threshold rule for such statistics by assuming a common alternative. A typical analysis process is involved in model selection and model prediction. Model selection aims to find a parsimonious, well-fitting model for the basic responses and model prediction uses summarizing statistics from the primary analysis to make statistical inferences [9]. However, starting with the summarizing statistics makes the model selection for the basic responses secondary and difficult, leading to inefficient tests [4]. In addition, assumptions about a common alternative may not always be feasible. For BH-FDR, the conventional t-statistics (and corresponding p-value) are used for testing the difference of means between two groups. The LR-FDR method models with the basic response, not summarizing statistics, which allows for different alternatives for each test. The likelihood approach provides an efficient way of controlling the FDR by simultaneously minimizing the FNDR and the useful information such as consistent estimates of effect size or proportion of null hypotheses.

In this study, we first applied the LR-FDR method to simulated data with extensive alternative proportion (hypometabolic areas in neuroimaging data) and then to brain positron emission tomography (PET) data of three groups: Alzheimer's disease (AD), questionable dementia (QD), and normal controls (NC). QD is also known as mild cognitive impairment (MCI), and the QD patients in our study were particularly at risk of developing dementia in the near future. We extended the model of Lee and Bjørnstad [4] to allow the distribution of test-statistic is asymmetric.

We compared the LR-FDR method to conventional thresholding using Benjamini and Hochberg's FDR (BH-FDR), to establish its efficiency when determining hypometabolic regions in AD and QD groups.

Methods

The LR-FDR method

Consider a hierarchical model for the basic responses. For the vth location within the brain and the jth individual in the control group ($v = 1, \ldots, N$ and $j = 1,\ldots, n_1$), suppose that the response y_{vj1} is modeled by

$$y_{vj1} = \xi_v + e_{vj1}, \tag{1}$$

where ξ_v is the mean parameter and $e_{vj1} \sim N(0, \phi_{v1})$. Then, the treatment (or disease) group has n_2 individuals ($j = 1, \ldots, n_2$), and the response y_{vj2} is modeled by

$$y_{vj2} = \xi_v + w_v + e_{vj2}, \tag{2}$$

where w_v is the treatment (or disease) effect, and $e_{vj2} \sim N(0, \phi_{v2})$. Thus, conditional on w_v, the difference between the means of the two groups,

$$dv = \bar{y}\, v2 - \bar{y}\, v1, \tag{3}$$

follows $N(w_v, \psi_v)$, with $\psi_v = \phi_{v1}/n_1 + \phi_{v2}/n_2$ and $\bar{y}\, v1 = \sum_j y_{vjk}/n_k$, for $v = 1, \ldots, N$ and $k = 1, 2$. To estimate $\psi = (\psi_1, \ldots, \psi_N)$, we use the unbiased estimators of ϕ_{vk} ($v = 1, \ldots, N$ and $k = 1, 2$),

$$\begin{aligned}\hat{\phi}_{vk} &= \frac{1}{n_k - 1}\sum_{j=1}^{n_k}\left(y_{vjk} - \bar{y}_{\,vk}\right)^2 \text{ and } \hat{\psi}_v \\ &= \hat{\phi}_{v1}/n_1 + \hat{\phi}_{v2}/n_2.\end{aligned} \tag{4}$$

To complete the model, specify the model for the treatment effects, w_v.

One-sided test

Let the null hypothesis H_v be the vth voxel is not abnormally activated (not different between two groups). Following Lee and Bjørnstad [4], we defined the binary random variable o_v, such that $o_v = 0$ if the null hypothesis H_v is true, $o_v = 1$ if H_v is false, and $p_s = P(o_v = s)$ for $s = 0$ or 1, with $p_0 + p_1 = 1$. Now, the multiple test problem can be viewed as predicting o_v.

Conditional on o_v, assume that w_v follows the normal distribution:

$$\text{Given } o_v = 0, w_v \sim N\left(0, \sigma^2\right),$$

$$\text{given } o_v = 1, w_v \sim N\left(\mu, \tau^2\right).$$

Here, we consider only normal distribution for w_v. However, this likelihood approach can be easily extended to other distributions. In this study, we prefer to have $\sigma^2 > 0$ since, typically, the null hypotheses "$w_v = 0$" are never *exactly true*, but rather $w_v = 0$, which can be

modeled by $0 < Var(w_v) = \sigma^2$. Here, ψ_v in (4) represents the within-test variation, σ^2 the between-test variation for Uninteresting cases, and τ^2 the between-test variation for the Interesting cases. If ψ_v is assumed to be common (i.e., $\psi_v = \psi$ for all v), this means that there is nothing special about any voxel in the alternative, and they are all statistically exchangeable. How can we determine whether voxels are all (statistically) exchangeable? Since we assume that the ψ_vs are not common and estimate them separately, we have a statistical model that allows all active voxels to have the same mean effect, but with different sampling variances. In addition to ϕ_{v1} and ϕ_{v2}, in this model, we have the fixed parameters $\theta = (p_0, \mu, \sigma^2, \tau^2)$.

We denote $d = (d_1, \ldots, d_N)^T$ and let w and o be the vectors of w_v and o_v, respectively. In this study, o is the inferential focus and w is a nuisance parameter which can be integrated out as follows:

$$\log f_\theta(d,o) = \log \int f_\theta(d,w,o)dw$$

$$= \sum_{v=1}^{N} [\log f_\theta(d_v|o_v) + \log f_\theta(o_v)],$$

where

$$\log f_\theta(d_v|o_v) = I(o_v=0)\left[-\frac{1}{2}\log\left(2\pi(\psi_v+\sigma^2)\right) - \frac{d_v^2}{2(\psi_v+\sigma^2)}\right] + I(o_v=1)\left[-\frac{1}{2}\log\left(2\pi(\psi_v+\tau^2)\right) - \frac{(d_v-\mu)^2}{2(\psi_v+\tau^2)}\right]$$

$$\log f_\theta(o_v) = I(o_v=0)\log p_0 + I(o_v=1)\log(1-p_0),$$

where $I(\cdot)$ is the indicator function.

To estimate the fixed parameters, θ, Lee and Bjørnstad [4] used the maximum-likelihood (ML) estimator for the log-likelihood,

$$l(\theta) = \sum_{v=1}^{N} \log f_\theta(d_v), \tag{5}$$

where $\log f_\theta(d_v) = \sum_{ov} \log f_\theta(d_v, o_v)$ and ψ_v are substituted by $\hat{\psi}_v$. This avoids the downward bias of the ML estimation owing to the large number of nuisance parameters, ψ_v, in the model [4].

Since $f_\theta(d_v, o_v)$ is a density function for a mixture, the unboundedness of likelihood might occur without a proper constraint on the parameters. However, Hathaway [10] pointed out that this problem can be resolved by a local maximizer of the likelihood in the interior of the parameter space that is consistent and asymptotically efficient. Therefore, to avoid the unboundedness problem, we are actually looking for a good local maximum of the likelihood, which would satisfy both $\hat{\sigma}^2 > 0$ and $\hat{\tau}^2 > 0$ [11]. To estimate θ, we used the expectation-maximization (EM) algorithm of Dempster, Laird and Rdin [12], with the proper initial values.

Let δ_v be a test for the vth null hypothesis, $H_v : \delta_v = 0$ (non-discovery) if H_v is not rejected, and $\delta_v = 1$ (discovery) if H_v is rejected. For some $\alpha > 0$, consider the loss function

$$L(\alpha; \delta_1, \ldots, \delta_N) = \sum_{v=1}^{N} \Big[I(o_v=1)I(\delta_v=0) + \alpha I(o_v=0)I(\delta_v=1)\Big]. \tag{6}$$

Lee and Bjørnstad [4] showed that the optimal decision rule, $\{\delta_1^\alpha, \ldots, \delta_N^\alpha\}$, that minimizes the risk with the loss function (6) is

$$\delta_v^\alpha = 1 \text{ if } R(d_v;\theta) > \alpha;$$

$$= 0 \text{ otherwise},$$

where $R(d_v;\theta) = \frac{P(o_v=1|d_v)}{P(o_v=0|d_v)} = \frac{p_1 N(d_v;\mu,\ \psi_v+\tau^2)}{p_0 N(d_v;0,\ \psi_v+\sigma^2)}$ is the likelihood ratio. Among tests with the common expected number of discoveries, this test is optimal in the sense that it controls the FDR with the smallest FNDR.

The outcomes of multiple tests can be summarized as in Table 1. Following Lee and Bjørnstad [4], we define the FDR and FNDR as $E(V)/E(D)$ and $E(N-N_0-S)/E(N-D)$, respectively. Benjamini and Hochberg defined the Fdr as $E(V/D)$, but Genovese and Wasserman [13] showed that $\mathrm{Fdr} = E(V/D)$ and $\mathrm{FDR} = E(V)/E(D)$ are asymptotically equivalent (in N) if the tests are independent. Suppose we want a test with an FDR level of κ. In this study, we first estimate the FDR as $\widehat{FDR}(\alpha)$, for each α. Then, we search for the cutoff α such that $\widehat{FDR}(\alpha) = \kappa$ to obtain the optimal test, $\{\delta_1^\alpha, \ldots, \delta_N^\alpha\}$, with an FDR control of κ. Lee and Bjørnstad [4] used the following estimator:

$$\widehat{FDR}(\alpha) = \frac{E(V)|_{\theta=\hat{\theta}}}{D},$$

which works well in their examples from genetic studies in which N is of the order of several thousands. However, in brain images for which $N = 329{,}694 >> 10{,}000$, we

Table 1 The outcomes of N multiple hypothesis tests

	Non-discovery	Discovery	Total
Null	$N_0 - V$	V	N_0
Alternative	$N - N_0 - S$	S	$N - N_0$
Total	$N - D$	D	N

found that D can sometimes be less than $E(V)\big|_{\theta=\hat{\theta}}$. To avoid this problem, we use the estimator

$$\widehat{FDR(\alpha)} = \frac{E(V)|_{\theta=\hat{\theta}}}{E(V)\big|_{\theta=\hat{\theta}} + E(S)\big|_{\theta=\hat{\theta}}}.$$

In our models, $\delta_\nu^\alpha = I(R(d_\nu;\theta) > \alpha) = I\left(d_\nu < c_\nu^l \text{ or } d_\nu > c_\nu^u\right)$. For a given α, the cutoff values c_ν^l and c_ν^u can be solved numerically from the equation $R(d_\nu;\theta) = \alpha$. Then,

$$\begin{aligned} E(V) &= \sum P\left(o_\nu = 0, \delta_\nu^\alpha = 1\right) = p_0 \sum P\left(\delta_\nu^\alpha = 1 | o_\nu = 0\right) \\ &= p_0 \sum \left[\Phi\left(c_\nu^l / \sqrt{\psi_\nu + \sigma^2}\right) + \tilde{\Phi}\left(c_\nu^u / \sqrt{\psi_\nu + \sigma^2}\right)\right], \\ E(S) &= \sum P\left(o_\nu = 1, \delta_\nu^\alpha = 1\right) = p_1 \sum P\left(\delta_\nu^\alpha = 1 | o_\nu = 1\right) \\ &= p_1 \sum \left[\Phi\left((c_\nu^l - \mu) / \sqrt{\psi_\nu + \sigma^2}\right)\right. \\ &\quad \left. + \tilde{\Phi}\left((c_\nu^u - \mu) / \sqrt{\psi_\nu + \sigma^2}\right)\right], \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and $\tilde{\Phi}(\cdot) = 1 - \Phi(\cdot)$. By plugging in the estimates of θ, we obtain an estimator for the FDR.

Two-sided test

In a two-sided problem, we may take only two actions. We can either simply accept or reject the null hypothesis without distinguishing between the positive and negative effects. In the case of brain data, it is important to assign abnormal regional changes at the voxel level. An abnormal voxel can be defined as an abnormally positive (hypermetabolic) or negative (hypometabolic) state. Especially positive activity might be associated with a treatment effect after treatment or functional compensation, while negative activity might be associated with functional deficit. This statistically specific determination of an abnormal voxel influences clinical interpretations. Therefore, this method allows us to consider the sign of the test statistic to decide whether the alternative discovery is positive or negative when the null hypothesis is rejected. In other words, we would never conclude that there is a discovery without stating whether the effect is positive or negative. As we show in the discussion on our results, in neuroimaging, an entire alternative can be either hypermetabolic or hypometabolic, whereas in genetics, alternatives often exist in both directions.

Extending the one-sided test model from the previous section, we used the *two-sided multiple testing with three actions* of Lee and Bjørnstad [4]. Here, the discrete random variable, o_ν, takes one of the three states: (a) $o_\nu = 0$ if the ith case is "Uninteresting;" (b) $o_\nu = 1$ if the ith case is "Interesting, with a Positive effect;" and (c) $o_\nu = -1$ if the ith case is "Interesting, with a Negative effect." In addition, $p_s = P(o_\nu = s)$ for s = −1, 0, 1, with $p_0 + p_1 + p_{-1} = 1$.

Consider the differences in (3). Suppose that, conditional on o_ν, w_ν follows a normal distribution, as follows:

$$\begin{aligned} &\text{Given } o_\nu = 0, w_\nu \sim N\left(0, \sigma^2\right), \\ &\text{given } o_\nu = 1, w_\nu \sim N\left(\mu_P, \tau_P^2\right), \\ &\text{given } o_\nu = -1,\ w_\nu \sim N\left(-\mu_N, \tau_N^2\right), \end{aligned}$$

where μ_P, $\mu_N > 0$. For simplicity of arguments, in this paper, we assume that $\mu_P = \mu_N = \mu$ and $\tau_P^2 = \tau_N^2 = \tau^2$. In this model, we have the fixed parameters $\theta = (p_0, p_1, \mu, \sigma^2, \tau^2)$, yielding a three-mixture model. Thus, we can use the EM algorithm to estimate θ. Since o_ν takes one of three states, the decision rule δ_ν also takes a value in {0, 1, −1}. In other words, $\delta_\nu = 0$ (non-discovery) if H_ν is not rejected, $\delta_\nu = 1$ if H_ν is rejected with a positive effect, and $\delta_\nu = -1$ if H_ν is rejected with a negative effect.

For some $\alpha_+ > 0$ and $\alpha_- > 0$, consider the loss function

$$\begin{aligned} L(\alpha_+, \alpha_-; \delta_1, \ldots, \delta_N) = \sum_{\nu=1}^{N} \Big[& I(o_\nu \neq 0) I(\delta_\nu \neq o_\nu) \\ & + I(o_\nu = 0)\{\alpha_+ I(\delta_\nu = 1) \\ & + \alpha_- I(\delta_\nu = -1)\} \Big] \end{aligned} \tag{7}$$

Then, the optimal decision rule that minimizes the risk in the loss function (7) is

$$\begin{aligned} \delta_\nu^{\alpha_+,\alpha_-} &= 1 \text{ if } R^+(d_\nu;\theta) > \alpha_+ \text{ and } R^+(d_\nu;\theta) - R^-(d_\nu;\theta) > \alpha_+ - \alpha_-; \\ &= -1 \text{ if } R^-(d_\nu;\theta) > \alpha_- \text{ and } R^+(d_\nu;\theta) - R^-(d_\nu;\theta) < \alpha_+ - \alpha_-; \\ &= 0 \text{ otherwise.} \end{aligned}$$

where

$$R^+(d_\nu;\theta) = \frac{P(o_\nu = 1|d)}{P(o_\nu = 0|d)} = \frac{p_1 N(d_\nu; \mu, \ \psi_\nu + \tau^2)}{p_0 N(d_\nu; 0, \ \psi_\nu + \sigma^2)},$$

$$R^-(d_\nu;\theta) = \frac{P(o_\nu = -1|d)}{P(o_\nu = 0|d)} = \frac{p_{-1} N(d_\nu; -\mu, \ \psi_\nu + \tau^2)}{p_0 N(d_\nu; 0, \ \psi_\nu + \sigma^2)}.$$

If we control the FDR at level κ for both directions using α_+ and α_-, the resulting two-sided test with three actions maintains the FDR at the same level. Furthermore, this optimal test allows more flexible analysis which can control the FDR at different level for each direction, for example, 0.05 for positive direction and 0.01 for negative direction. In fact, Lee and Bjørnstad [4] showed that the resulting multiple two-sided test with three actions $\left\{\delta_1^{\alpha_+,\alpha_-}, \ldots, \delta_N^{\alpha_+,\alpha_-}\right\}$ minimizes the expected number of false assignments.

Simulation data

Simulation data were generated with a dimension of 400×400 pixels. We set the proportion of positive pixels to 80% (Simulation I) and 60% (Simulation II) per 160,000 total pixels, considering that the estimates of p_1 were high in our PET data. For each simulation setting, we varied $\mu = 1, 3, 5$ and fixed $\sigma^2 = 0.3$ and $\tau^2 = 0.5$. From (1) and (2), we randomly generated y_{vj1} and y_{vj2} for $v = 1, \ldots, 160,000$, $j_1 = 1, \ldots, 30$, and $j_2 = 1, \ldots, 30$. For each simulation, we generated 100 simulation data sets and applied both the LR-FDR and BH-FDR methods to control the FDR at the 0.05 level.

AD and QD PET data

PET data were composed of two types of patient groups and one control group. The first group consisted of 22 probable AD patients (mean age, 66.9 ± 7.2) with moderate dementia according to the criteria of the Mini-Mental State Examination (MMSE), with a mean MMSE score of 13 ± 5.0, and a Clinical Dementia Rating (CDR) score between 1 and 3. Generally, the MMSE score can be indicated severe (<9), moderate (10–18), mild (19–24) cognitive impairment. The AD patients suffered progressive memory loss, but had no disturbance of consciousness. The second group comprised 24 QD patients (mean age, 67.3 ± 9.0) who showed objective evidence of memory and/or cognitive impairments, but did not satisfy the criteria for AD. Their CDR scores were all 0.5, and their mean MMSE scores were 23 ± 4.1.

All the patients were diagnosed by clinical evaluation using the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association AD criteria as a guideline. The two patient groups described above were compared with 44 normal control (NC) subjects (mean age, 68.9 ± 5.2). These NC subjects were recruited from the Health Care Center at Seoul National University Hospital and had no history of neurological disorders, psychiatric disorders, significant medical conditions, or substance abuse. For subject screening, the Korean version of the modified MMSE and the Mood Evaluation Scale were used, and only right-handed subjects were included in the study. Furthermore, there was no significant age difference among the three groups. This study was approved by the institutional review board (IRB) of the Seoul National University Hospital. PET data of our patients were only part of the patient's standard care. We used patient's data from database obtained from 1996 to 1999. Normal controls were recruited for other study purpose (i.e., creation of Korean Standard Brain Template) from Center for Health Promotion and Optimal Aging of Seoul National University Hospital in 2001 [14], who provided informed consent which was verbal form. For our research using identifiable human data, such as PET data in database of department, although we didn't receive documented informed consent from participants, IRB of our institute decided that this study protocol was applicable to exceptional situations where consent would be impracticable to obtain due to reuse storage data in database. Also our study was conducted in a manner that minimizes possible abuse to human subject's health and rights and no clinical intervention was performed for our study.

PET image acquisition

^{18}F-FDG PET images were obtained using an ECAT EXACT 47 (Siemens-CTI, Knoxville, TN, USA) PET scanner with an intrinsic resolution of 5.2 mm FWHM. After obtaining a transmission scan measured by ^{68}Ge rod sources for attenuation correction, an emission scan was obtained. During the resting state, ^{18}F-FDG was administered in doses of 370 MBq (10 mCi) for 30 min to obtain a static emission scan. All participants were scanned under the normal environmental noise conditions in the scanner room. For transaxial image reconstruction a filtered back-projection algorithm (Shepp-Logan filter at a cutoff frequency of 0.3 cycles/pixel as $128 \times 128 \times 47$ matrices of size $2.1 \times 2.1 \times 3.4$ mm) was used.

Image processing

All PET images were preprocessed using Statistical Parametric Mapping (SPM 2, University College of London, UK) and implemented in the Matlab 6.5 (Mathworks Inc., USA) environment. After spatial normalization to the Montreal Neurological Institute (MNI) space, all images were smoothed with a Gaussian filter of 16 mm full width at half maximum (FWHM). The PET signal intensity was normalized to the individual's total mean count for the cerebellum. This region was chosen as a reference region because it remains relatively unaffected until late in the progression of AD, if at all. To remove non-brain voxels, normalized and smoothed PET images were exclusively masked with a binary brain mask image. The same masked PET images were applied to both LR-FDR and BH-FDR methods using R software.

Results

Simulation results

We applied the proposed LR-FDR method to the simulated data set. The simulated data had a pixel dimension of 400×400, which yielded a total 160,000 tests. We considered two simulation settings with varying p_1, the proportion of pixels with $o_v = 1$: p_1 was 80% in Simulation I and 60% in Simulation II. Figure 1 shows the FDR and FNDR results based on the 100 simulated data sets.

The LR-FDR method yielded a smaller FNDR than the BH-FDR method (Figure 1): 20% lower when $\mu = 3$ or 5 in Simulation I ($p_1 = 80\%$) and 5% lower when $\mu = 3$ in

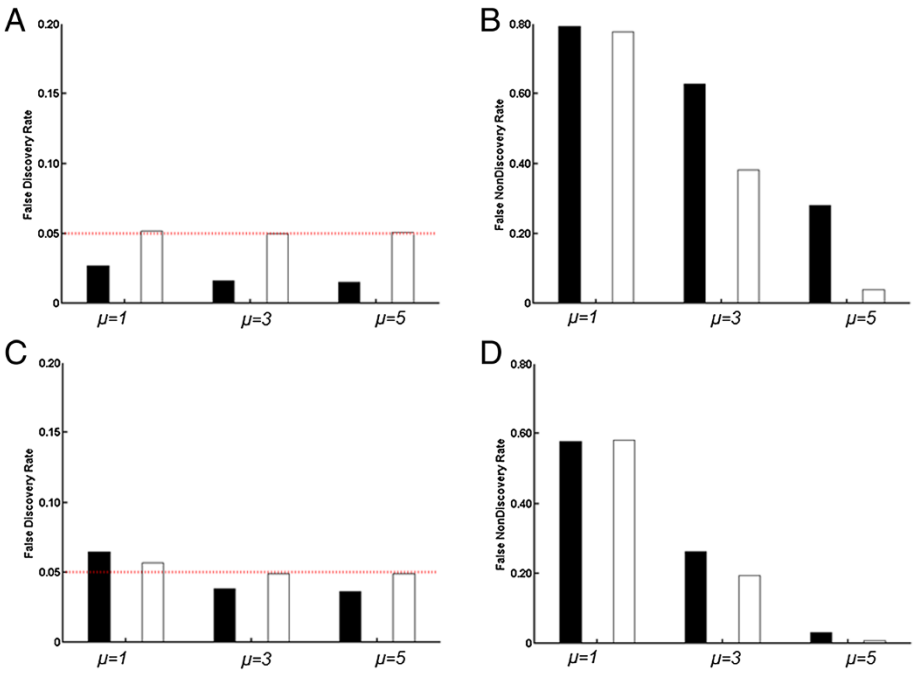

Figure 1 The averaged FDR and FNDR. Within each panel, black and white bars represent the BH-FDR and LR-FDR methods, respectively. The alternative proportions of data are 80% and 60% in Simulation I **(A and B)** and II **(C and D)**, respectively. In each simulation setting, depending on μ, three parameter settings are presented.

Simulation II ($p_1 = 60\%$). The LR-FDR method yielded FDR results quite close to 0.05 in both simulation settings. The minimum and maximum of the average FDR from the 100 repeated tests were 0.049 and 0.056, respectively, across all settings. The BH method often yielded a more conservative FDR control for most of the settings.

Results of the AD and QD data analysis

In probable AD cases, all methods (one-sided test of LR-FDR, two-sided test of LR-FDR, conventional BH-FDR methods) revealed hypometabolic regions at FDR level 0.01 (Figure 2). Both the one-sided and two-sided tests of LR-FDR showed hypometabolism in the bilateral posterior cingulate, frontal, temporal, and parietal areas, the extent of which was wider than that shown by conventional BH-FDR. More specifically, the LR-FDR method showed that the hypometabolic regions spread to the posterior prefrontal and anterior occipital regions in the AD group. No hypometabolic areas were observed in the sensorimotor and visual areas by any of the methods. Quantification using the LR-FDR method generally found a greater number of voxels than did the BH-FDR method (Table 2). In the QD cases, the LR-FDR method showed hypometabolic regions in both medial temporal areas, including the hippocampus and anterior frontal cortex (Figure 3). The hypometabolic voxels in the medial temporal regions were found more easily using LR-FDR method at 0.05, 0.01, 0.005, and 0.001. However, no hypometabolic region was found by BH-FDR method with controlling FDR at 0.01 (Table 3).

The estimates of the fixed parameters are shown in Table 4. In the AD cases, two-sided tests give the effect size, $\hat{\mu} = 4.524$, and the estimated probability of "Interesting, with a Negative effect," $\hat{p}_{-1} = 0.771$. Since $\hat{p}_1$ approaches 0 in the two-sided test, both tests have the same parameter estimates and the same number of significant voxels. In the QD cases, very few hypermetabolic region was found.

The distribution of $t_v = d_v/\sqrt{\psi_v + \sigma^2}$ at the null $o_v = 0$ was $N(0, 1)$. However, Figure 4 shows that most t_v -values for both the AD and QD groups were located on the left of the theoretical null distribution, $N(0, 1)$. Lee and Bjørnstad [4], when analyzing genetic data, assumed that $p_1 = p_{-1}$, but we did not do so here, as in our neuroimaging data, $p_1 \ll p_{-1}$. For both the AD and QD PET data, the symmetric model (with $p_1 = p_{-1}$) was not plausible. Therefore, we avoided using the wrong symmetric model by estimating, p_1 and p_{-1} separately. To check the goodness of fit, we first generated a synthetic sample, d_v^*, from the fitted model, $f_\theta(d_v)$, using the estimated parameters in Table 4. Figure 4 shows the histogram of d_v^*. Since the shapes of the histograms of the d_v (from the real data) and

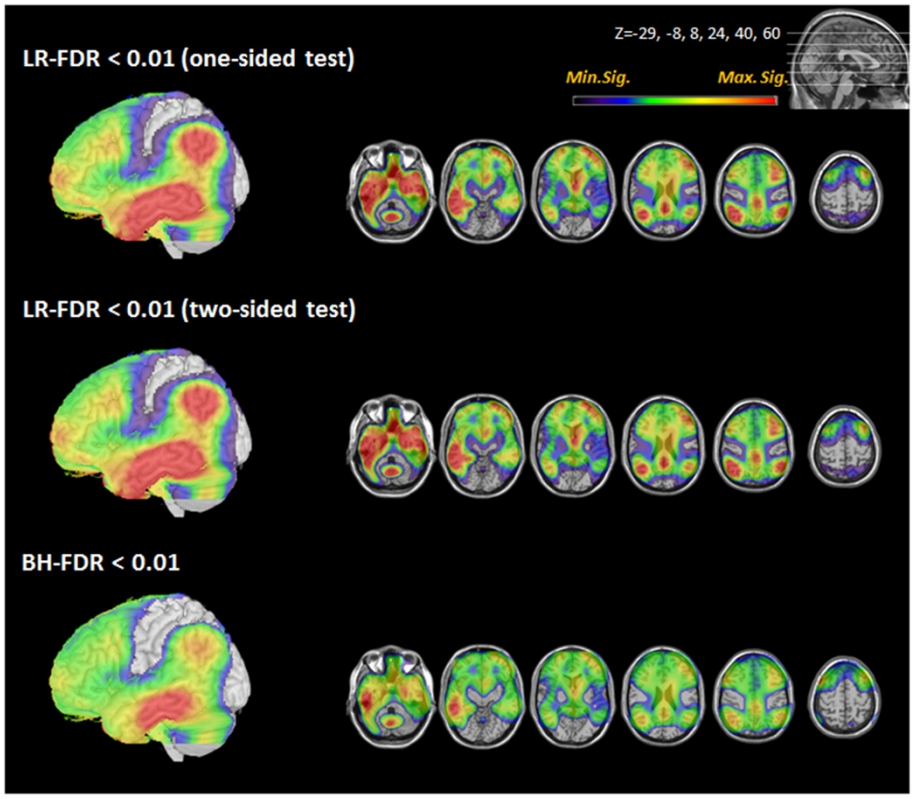

Figure 2 Brain regions with significantly lower FDG uptake in probable AD compared to NC. Regions with lower FDG uptake in probable AD are displayed. The left hemisphere is shown as a 3D volume rendering. The reduction in the FDG uptake in the temporal, parietal, and posterior prefrontal regions was commonly found in the LR-FDR and BH-FDR methods. Extensive hypometabolic areas extending to posterior prefrontal were detected with the LR-FDR one-sided and two-sided tests. The color bar range from minimum to maximum significance level denotes the significance of the likelihood ratio in both LR-FDR methods and of the *p*-value in BH-FDR method. (AD: Alzheimer's disease; FDR: False discovery rate; LR-FDR: Likelihood ratio false discovery rate; NC: Normal controls).

d_v^* (from the synthetic data) were similar, we could say that resulting model fitting was appropriate.

In AD group, the result of the LR-FDR one-sided test was the same as that of the LR-FDR two-sided test with three actions, because in these data, there was no positive effect ($\hat{p}_1 = 0$). In other words, no hypermetabolic region was found in AD patient group.

Table 2 Total number of voxels in the whole brain with significant hypometabolism at different threshold levels

Comparisons	Threshold levels	Number of significant voxels		
		LR-FDR (one-sided)	LR-FDR (two-sided)	BH-FDR
NC > AD	FDR 0.001	194789	194789	165451
	FDR 0.005	213881	213881	207426
	FDR 0.01	223091	223091	221507
	FDR 0.05	249706	249706	251100
NC > QD	FDR 0.001	8471	7740	47
	FDR 0.005	18094	16767	140
	FDR 0.01	25624	23813	212
	FDR 0.05	53229	50287	22038

AD: Alzheimer's disease; NC: Normal controls; QD: Questionable dementia.

Discussion

In this study, we applied the LR-FDR method to neuroimaging data. We found that the LR-FDR method increased the detection capability in the simulated as well as brain PET data, allowing us to decrease the FNDR and find larger areas of abnormality under the given level of the FDR, respectively. Decreasing the FNDR worked when the difference of the means of the two groups was within a range specified in the simulation study. When we compared the two patient groups (AD and QD) with NC group, the three actions of 1, 0, and −1, corresponding to positive (normal < patients), null (normal = patients), and negative (normal > patients) differences, revealed areas of hyper-, eu-, and hypo-metabolism, respectively. Only negative results (i.e., hypometabolism) in AD patients as compared to normal were obtained and visualized in both the one-sided test and the two-sided test with three actions. In the two-sided test with three actions,

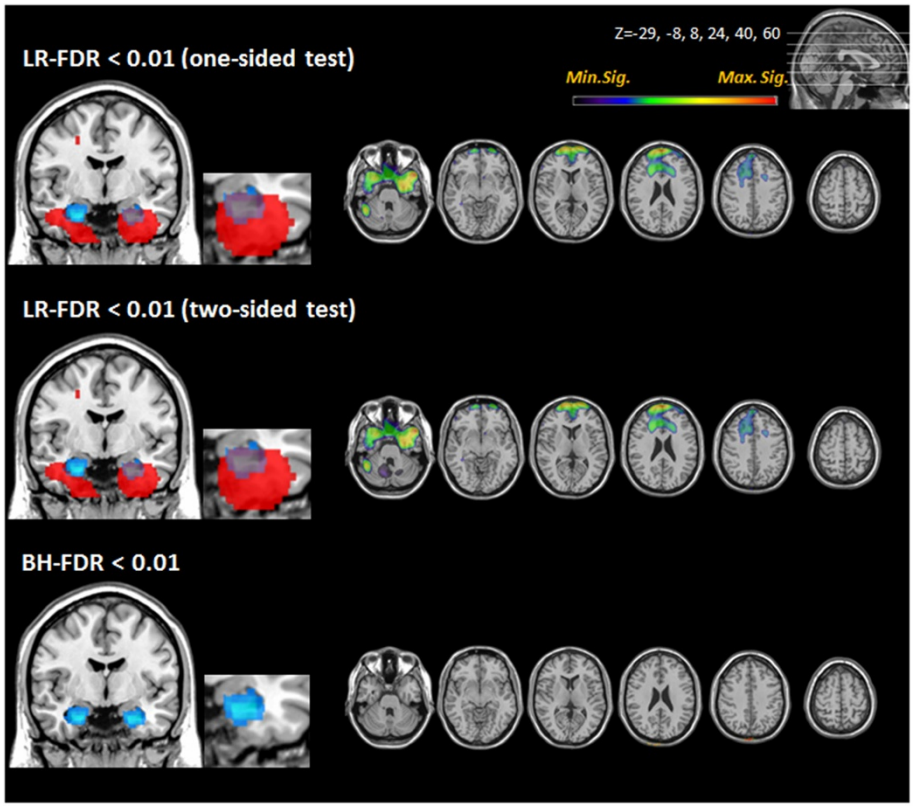

Figure 3 Brain regions with significantly lower FDG uptake in QD compared to NC. The coronal view in the left column shows hypometabolism in the medial temporal regions in QD. An anatomical map of the hippocampus is displayed in blue. Regions with a lower FDG uptake are displayed in red. The LR-FDR one-sided and two-sided tests disclosed more extensive hypometabolic areas in both temporal lobes than did the BH-FDR method. (FDR: False discovery rate; LR-FDR: Likelihood ratio false discovery rate; NC: Normal controls; QD: Questionable dementia).

the estimated probability of a hypermetabolic region was zero in cases with AD. In these cases, extensive regional metabolic reduction was found throughout the brain, with the same degree, by the one-sided test and two-sided test with three actions.

In existing literature, several reports have stated that abnormalities in glucose metabolism are probably present in the medial part of the temporal lobes early in the development of AD [15,16]. QD or MCI subjects (CDR score of 0.5) are likely to show initial minor abnormalities [17,18] that, in some cases, have progressed to probable AD upon follow-up [16,19,20]. Various investigators have attempted to find the predictive areas of abnormality, by using both FDG PET [16] and MRI using voxel-based morphometry [21], to predict future development of AD. Among these predictors are medial temporal lobe involvement of MRI signal loss (atrophy) [18-21], accumulation of neutritic plaques [22], and hypometabolism [16,23].

The expansion of hypometabolism to the temporal, cingulate, or other cortices was a common finding in AD. However, the right/left asymmetry of involvement or the exact nature of the abnormality in the hippocampus shown by FDG PET or MRI has not been consistently

Table 3 Total number of voxels in hippocampus with significant hypometabolism at different threshold levels

Comparisons	Threshold levels	Number of significant voxels		
		LR-FDR one-sided (L/R)	LR-FDR two-sided (L/R)	BH-FDR (L/R)
NC > QD	FDR 0.001	0/117	0/117	0/0
	FDR 0.005	1/181	1/184	0/0
	FDR 0.01	12/217	12/217	0/0
	FDR 0.05	118/303	121/303	0/158

NC: Normal controls; QD: Questionable dementia.

Table 4 Parameter estimates

Comparisons	Tests	μ	σ^2	τ^2	P_0	P_1	P_{-1}
NC > AD	One-sided	−4.524	0.001	9.437	0.229	0.771	-
	Two-sided	4.524	0.001	9.437	0.229	0.000	0.771
NC > QD	One-sided	−1.437	0.000	0.696	0.669	0.331	-
	Two-sided	1.479	0.000	0.594	0.674	0.004	0.323

AD: Alzheimer's disease; NC: Normal controls; QD: Questionable dementia.

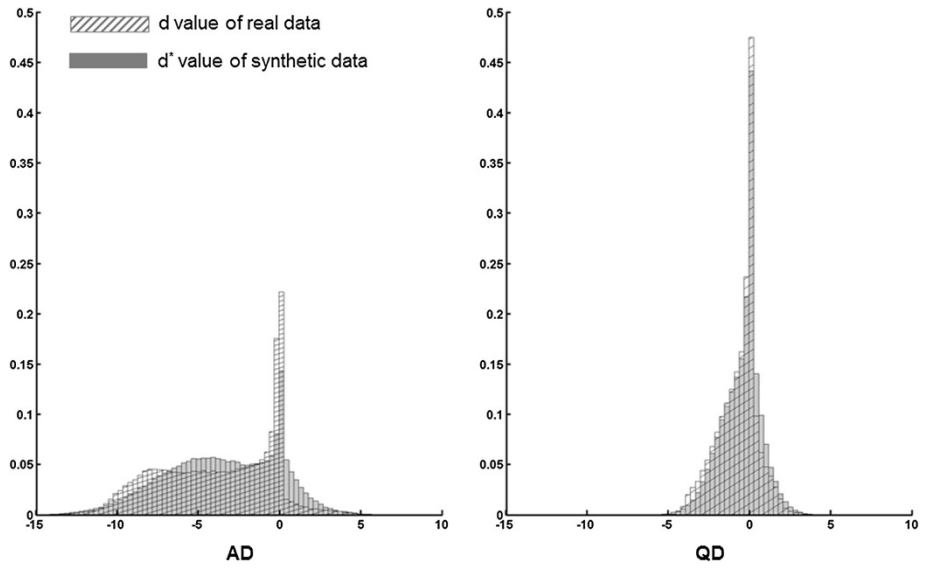

Figure 4 Histograms of real and synthetic data. Histograms of d_v^* of the generated synthetic data from fitted model (gray histogram) and d_v of real data (hatched histogram) (left: AD cases, right: QD cases) (AD: Alzheimer's disease; QD: Questionable dementia).

reported [18,24-26]. This might be due to differences between patients and the normal populations examined, but might also be due to differences in the statistical methods used to detect abnormalities. Thus far, considerable effort has been made to control false positives, but less effort has gone into minimizing false negatives. Using this novel LR-FDR method to minimize the FNDR, we found right-dominant abnormalities in the hippocampus in a relatively small patient group.

Using simulation studies, we showed that the LR-FDR method controlled the FDR quite near to the stated level. In contrast, the BH-FDR method did not maintain the stated FDR level, and instead became more conservative (i.e., a lower FDR then set beforehand). Furthermore, the LR-FDR method reduced the FNDR significantly in certain situations, according to the simulation study, as compared to the BH-FDR method. The FNDR reduction became greater as p_1 increased. In the neuroimaging data, such as the AD PET data, p_1 was larger (e.g., 0.771), whereas in the genomic studies, p_1 was often small (i.e., less than 0.05). The BH-FDR method assumes that $\sigma^2 = 0$. However, the LR-FDR method allows for non-zero between-test variations ($\sigma^2 > 0$ or $\tau^2 > 0$). In our PET study using real imaging data, we found that the maximum likelihood estimates of τ^2 are very different from zero. In a neuroimaging data analysis, the LR-FDR method was preferred over the BH-FDR method. The LR-FDR method had a higher detection capability, and showed extensive hypometabolic regions in patients with AD or QD. Especially, in QD group, no significant area was found in BH-FDR at 0.01, although the hypometabolic voxels were 229 in LR-FDR method. One possibility is that the hippocampal region was falsely assigned as a null in the BH-FDR method at this threshold level.

The data used in this study were drawn from Lee, Kang, Jang, Cho, Kang, Lee, Kang, Lee, Woo and Lee [27], and the assessment of cerebral glucose metabolism by FDG-PET in a resting state correlated well with the progression of disease severity in patients with AD [23,28]. Unlike patients with cognitive deterioration associated with old age, patients with AD showed decreased FDG uptake in both parietal regions, including the posterior cingulate and temporal areas and the frontal cortices, as the disease progressed [29,30]. Primary sensory and motor cortices, as well as visual and deep gray cortices remained relatively intact in AD until late in the disease progression [31]. FDG-PET results, analyzed by all three methods, showed a characteristic spatial pattern of glucose hypometabolism in the parietal, temporal, and posterior prefrontal regions in patients with AD, as compared to the NC group. In AD cases, the pattern of distribution was similar. However, unlike the conventional methods, the LR-FDR method showed more extensive hypometabolic areas, extending symmetrically to posterior prefrontal cortices.

Hippocampal atrophy was once thought to be a discriminant feature in individuals with MCI at risk of AD [18,21]. In our investigation, the LR-FDR method could disclose that reduced FDG uptake in the hippocampal region is a discriminator between normal and QD patients [32,33]. The BH-FDR method showed no temporal hypometabolic result. In contrast, the LR-FDR method

revealed hypometabolism in bilateral medial temporal areas. The hypometabolism seen on the right side was more extensive and severe in LR-FDR method.

We showed that the LR-FDR method for two-sided multiple testing with three actions can be applied to neuroimaging data analysis to find hypermetabolic or hypometabolic regions. In the search for a pre-symptomatic imaging biomarker in the prodromal phase of AD (i.e., QD), we propose that the LR-FDR method is the most efficient tool and, therefore, optimizes the chances for success. According to the good fitting of the model shown in Figure 4, we could say that the non-symmetric model fitting and efficient analysis was feasible to yield robust results from the LR-FDR method, using either the one-sided test or two-sided test with three actions. In the non-symmetric cases, none of the methods employed by Lee and Bjørnstad [4] worked, assuming $p_1 = p_{-1}$. This is the advantage our LR-FDR method holds, when applied to neuroimaging data, over any existing p-value based methods.

The extended likelihood principle of Bjørnstad [5] means that if the assumed model is correct, all information on the unknowns is in the extended likelihood. Therefore, this can be the basis for the most efficient test. However, if the assumed model is not correct, the likelihood method may fail. All the existing multiple testing procedures have been developed without considering a proper model choice, so that, as Lee and Bjørnstad [4] showed, existing methods may not maintain the stated FDR level if any of their model assumptions are wrong. Under the likelihood approach, we can use the likelihood-based model-checking and model-selection procedures to enhance the performance of the test [4].

After reviewing the simulation data, we were surprised that the BH-FDR and LR-FDR methods produced so high an FNDR when the difference was small, for example, $\mu = 1$. We need to improve the methods to obtain robust results, even when the alternative and null distributions overlap by so much. Another interesting area of future research would be to study robust models for various violations of model assumptions using double hierarchical generalized linear models [7,34]. Furthermore, the neuroimaging data are actually spatially correlated among the voxels. Owing to the difficulty in specifying the full spatial dependency, we assumed independence over voxels. Genovese, Roeder and Wasserman [35] showed that exploiting the dependency structure improved the power. Thus, a further extension of the LR-FDR method to a spatially correlated model would be a promising prospect for future work.

Conclusions

We applied the LR-FDR method to PET data from AD and QD patients and compared the performance to that of conventional BH-FDR method. We found that the LR-FDR method enabled us to find more voxels with a congruent distribution. Based on our findings from the AD and QD PET subjects and our simulation study, proving the increased efficiency, bilateral hippocampal hypometabolism might serve as a marker for QD. It would be interesting to extend this approach to perform individual analyses of PET or MRI images to find a meaningful region of brain. A prospective study of a cohort of subjects with QD (or MCI), in which individuals might show a conversion to AD, is warranted, and the LR-FDR method would prove advantageous in such studies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DL analysed the data, contributed analysis tool and drafted the manuscript. HK analysed the data, drafted the manuscript and interpreted the results. EK analysed the data and visualized the results. HL helped to set and analyse simulation data. HJK and YK collected data and participated in design and coordination. YL developed the analysis tool, conceived and designed the analysis, and drafted the manuscript. DSL conceived and designed the analysis, interpreted the results and drafted manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (MEST) (grant No. 2011–0030810, 2011–0030815 and 2014M3C7A1062896).

Author details

[1]Department of Statistics, Ewha Womans University, Seoul, Korea. [2]Department of Nuclear Medicine, Seoul National University College of Medicine, Seoul, Korea. [3]Data Science for Knowledge Creation Research Center, Seoul National University, Seoul, Korea. [4]Interdisciplinary Program in Cognitive Science, Seoul National University, Seoul, Korea. [5]Department of Nuclear Medicine, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea. [6]Department of Statistics, Seoul National University, Seoul, Korea. [7]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, and College of Medicine, Seoul National University, Seoul, Korea.

Received: 18 September 2014 Accepted: 15 January 2015
Published: 30 January 2015

References

1. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57(1):289–300.
2. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage. 2002;15(4):870–8.
3. Cohen A, Sackrowitz H. More on the inadmissibility of step-up. J Multiv Anal. 2007;98:481–92.
4. Lee Y, Bjørnstad JF. Extended likelihood approach to large-scale multiple testing. J Roy Stat Soc B. 2013;75(3):553–75.
5. Bjørnstad JF. On the generalization of the likelihood function and likelihood principle. J Am Stat Assoc. 1996;91:791–806.
6. Lee Y, Nelder JA. Hierarchical generalized linear models (with discussion). J Roy Stat Soc B. 1996;58:619–78.
7. Lee Y, Nelder JA, Pawitan Y. Generalized linear models with random effects : unified analysis via h-likelihood. Boca Raton, FL: Chapman & Hall/CRC; 2006.
8. Efron B. The Future of Indirect Evidence. Stat Sci. 2010;25(2):145–57.
9. McCullagh P, Nelder JA. Generalized linear models 2nd ed. London; New York: Chapman and Hall; 1989.
10. Hathaway RJ. A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. Ann Stat. 1985;13(2):795–800.

11. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction 2nd ed. New York: Springer; 2009.
12. Dempster A, Laird N, Rdin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Statist Soci B. 1977;39:1–38.
13. Genovese CR, Wasserman L. Operating characteristics and extensions of the FDR procedure. J Roy Stat Soc B. 2002;64:499–518.
14. Lee JS, Lee DS, Kim J, Kim YK, Kang E, Kang H, et al. Development of Korean standard brain templates. J Kor Med Sci. 2005;20(3):483–8.
15. De Santi S, de Leon MJ, Rusinek H, Convit A, Tarshish CY, Roche A, et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiol Aging. 2001;22(4):529–39.
16. Morbelli S, Piccardo A, Villavecchia G, Dessi B, Brugnolo A, Piccini A, et al. Mapping brain morphological and functional conversion patterns in amnestic MCI: a voxel-based MRI and FDG-PET study. Eur J Nucl Med Mol Imaging. 2010;37(1):36–45.
17. Almkvist O, Basun H, Backman L, Herlitz A, Lannfelt L, Small B, et al. Mild cognitive impairment–an early stage of Alzheimer's disease? J Neural Transm Suppl. 1998;54:21–9.
18. Wolf H, Grunwald M, Kruggel F, Riedel-Heller SG, Angerhofer S, Hojjatoleslami A, et al. Hippocampal volume discriminates between normal cognition; questionable and mild dementia in the elderly. Neurobiol Aging. 2001;22(2):177–86.
19. Chetelat G, Fouquet M, Kalpouzos G, Denghien I, De la Sayette V, Viader F, et al. Three-dimensional surface mapping of hippocampal atrophy progression from MCI to AD and over normal aging as assessed using voxel-based morphometry. Neuropsychologia. 2008;46(6):1721–31.
20. Chetelat G, Landeau B, Eustache F, Mezenge F, Viader F, de la Sayette V, et al. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. Neuroimage. 2005;27(4):934–46.
21. Risacher SL, Saykin AJ, West JD, Shen L, Firpi HA, McDonald BC. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr Alzheimer Res. 2009;6(4):347–61.
22. Koivunen J, Scheinin N, Virta JR, Aalto S, Vahlberg T, Nagren K, et al. Amyloid PET imaging in patients with mild cognitive impairment: a 2-year follow-up study. Neurology. 2011;76(12):1085–90.
23. Silverman DH, Small GW, Chang CY, Lu CS, Kung De Aburto MA, Chen W, et al. Positron emission tomography in evaluation of dementia: Regional brain metabolism and long-term outcome. JAMA. 2001;286(17):2120–7.
24. Apostolova LG, Dinov ID, Dutton RA, Hayashi KM, Toga AW, Cummings JL, et al. 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. Brain. 2006;129(Pt 11):2867–73.
25. Geroldi C, Laakso MP, DeCarli C, Beltramello A, Bianchetti A, Soininen H, et al. Apolipoprotein E genotype and hippocampal asymmetry in Alzheimer's disease: a volumetric MRI study. J Neurol Neurosur Psych. 2000;68(1):93–6.
26. Tapiola T, Pennanen C, Tapiola M, Tervo S, Kivipelto M, Hanninen T, et al. MRI of hippocampus and entorhinal cortex in mild cognitive impairment: a follow-up study. Neurobiol Aging. 2008;29(1):31–8.
27. Lee DS, Kang H, Jang MJ, Cho SS, Kang WJ, Lee JS, et al. Application of false discovery rate control in the assessment of decrease of FDG uptake in early Alzheimer dementia. Korean J Nucl Med. 2003;37(6):374–81.
28. Desgranges B, Baron JC, Lalevee C, Giffard B, Viader F, de La Sayette V, et al. The neural substrates of episodic memory impairment in Alzheimer's disease as revealed by FDG-PET: relationship to degree of deterioration. Brain. 2002;125(Pt 5):1116–24.
29. Alexander GE, Chen K, Pietrini P, Rapoport SI, Reiman EM. Longitudinal PET Evaluation of Cerebral Metabolic Decline in Dementia: A Potential Outcome Measure in Alzheimer's Disease Treatment Studies. Am J Psychiatry. 2002;159(5):738–45.
30. Langbaum JB, Chen K, Lee W, Reschke C, Bandy D, Fleisher AS, et al. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Neuroimage. 2009;45(4):1107–16.
31. Frisoni GB, Pievani M, Testa C, Sabattoli F, Bresciani L, Bonetti M, et al. The topography of grey matter involvement in early and late onset Alzheimer's disease. Brain. 2007;130(Pt 3):720–30.
32. Li Y, Rinne JO, Mosconi L, Pirraglia E, Rusinek H, DeSanti S, et al. Regional analysis of FDG and PIB-PET images in normal aging, mild cognitive impairment, and Alzheimer's disease. Eur J Nucl Med Mol Imaging. 2008;35(12):2169–81.
33. Mosconi L, Tsui WH, De Santi S, Li J, Rusinek H, Convit A, et al. Reduced hippocampal metabolism in MCI and AD: automated FDG-PET image analysis. Neurology. 2005;64(11):1860–7.
34. Lee Y, Nelder JA. Double hierarchical generalized linear models (with discussion). Appl Statis. 2006;55:139–85.
35. Genovese CR, Roeder K, Wasserman L. False discovery control with p value weighting. Biometrika. 2006;93:509–24.

doi:10.1186/1471-2288-15-9
Cite this article as: Lee *et al.*: Optimal likelihood-ratio multiple testing with application to Alzheimer's disease and questionable dementia. *BMC Medical Research Methodology* 2015 **15**:9.