Presenting the Uncertainties of Odds Ratios Using Empirical-Bayes Prediction Intervals

Wan-Yu Lin^{1,2}, Wen-Chung Lee^{1,3*}

1 Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, **2** Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, United States of America, **3** Research Center for Genes, Environment and Human Health, National Taiwan University, Taipei, Taiwan

Abstract

Quantifying exposure-disease associations is a central issue in epidemiology. Researchers of a study often present an odds ratio (or a logarithm of odds ratio, logOR) estimate together with its confidence interval (CI), for each exposure they examined. Here the authors advocate using the empirical-Bayes-based 'prediction intervals' (PIs) to bound the uncertainty of logORs. The PI approach is applicable to a panel of factors believed to be exchangeable (no extra information, other than the data itself, is available to distinguish some logORs from the others). The authors demonstrate its use in a genetic epidemiological study on age-related macular degeneration (AMD). The proposed PIs can enjoy straightforward probabilistic interpretations—a 95% PI has a probability of 0.95 to encompass the true value, and the expected number of true values that are being encompassed is $0.95m$ for a total of m 95% PIs. The PI approach is theoretically more efficient (producing shorter intervals) than the traditional CI approach. In the AMD data, the average efficiency gain is 51.2%. The PI approach is advocated to present the uncertainties of many logORs in a study, for its straightforward probabilistic interpretations and higher efficiency while maintaining the nominal coverage probability.

Citation: Lin W-Y, Lee W-C (2012) Presenting the Uncertainties of Odds Ratios Using Empirical-Bayes Prediction Intervals. PLoS ONE 7(2): e32022. doi:10.1371/journal.pone.0032022

Editor: Sten H. Vermund, Vanderbilt University, United States of America

Received: October 8, 2011; **Accepted:** January 18, 2012; **Published:** February 21, 2012

Copyright: © 2012 Lin and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was supported by National Science Councils, Taiwan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wenchung@ntu.edu.tw

Introduction

Quantifying the association between an exposure and a disease is a central issue in epidemiology. Although it has some limitations [1,2], odds ratio (OR) is probably the most widely used measure of exposure-disease association in epidemiology. For a dichotomous exposure, OR is estimated by the famous formula: $OR = \frac{ad}{bc}$, where a is the number of exposed cases; b , exposed controls; c , unexposed cases; and d , unexposed controls. To estimate the OR for an exposure with scale beyond simple yes/no, and simultaneously for multiple exposures, epidemiologists will need more than pencil and paper, but rather a logistic regression model to fit their data. Such a model fitting is usually performed by using statistical packages. The coefficients (except the intercept term) in a logistic regression model are the maximum likelihood estimates (MLEs) of the natural logarithms of odds ratios (logORs).

The ORs obtained from a study with a finite sample size are naturally subject to random variation to some degree, and therefore should not be taken as the true parameter values in and of themselves. To acknowledge this, epidemiologists often place a confidence interval (CI) around an OR estimate for each and every OR in the study. For a dichotomous exposure, the asymptotic 95% CI of logOR is simply: $\log\left(\frac{ad}{bc}\right) \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$. (The formula of the exact CI of a logOR can be found in Agresti [3].) For more general situations, the task of calculating CIs is again charged upon the logistic regression model.

Most epidemiologic studies are designed to simultaneously evaluate multiple risk factors. For example, with the advent of high-throughput technologies, genetic epidemiologists are often confronted with a large number of single nucleotide polymorphisms (SNPs), each calling for an OR (or a logOR) estimate and a CI. The total number of SNPs in a study can be in hundreds, thousands or even millions. Calculating CIs for so many logORs poses no special challenge. However, determining how to correctly interpret them often troubles even an experienced epidemiologist. Does the 'confidence level' associated with a CI signify the same meaning as 'chance' or 'likelihood' in our everyday usages, such that a 95% CI for a particular logOR has a probability of 0.95 to encompass the true logOR value? Unfortunately it does not. The confidence level of a CI actually relies on a 'repeated-sampling' interpretation. For a particular 95% CI, this means that if the procedure of producing it were to be repeated on multiple samples, the calculated 95% CIs (which would differ in location and in length for each sample) would encompass the true logOR value for 95% of the time [4]. This interpretation is based on sampling properties that are seldom realized in epidemiologic studies [2]. In fact, researchers in a study only get to work on one sample, the one in their study. It is cumbersome to have to conjure up multiple imaginary samples just to interpret a CI.

In this paper we advocate using the 'prediction intervals' (PIs) instead, to bound the uncertainty of the many logORs estimated in a study. The PI approach is applicable to a panel of factors believed to be exchangeable (that is, no extra information, other

than the data itself, is available to distinguish some logORs from the others). For example, when we evaluate the associations of a panel of SNPs in a chromosomal region with a disease outcome, we may not distinguish certain SNPs from the others before seeing the data we have. Like conventional CIs, the calculation of PIs presents no special difficulty, involving just one extra step of simple arithmetic operations of the output of a logistic regression model. But unlike CIs, the proposed PIs can enjoy straightforward probabilistic interpretations; a 95% PI has a probability of 0.95 to encompass the true value, and the expected number of true values that are being encompassed is $0.95m$ for a total of m 95% PIs. Furthermore, we will show that the proposed PI approach is theoretically more efficient (producing shorter intervals) than the traditional CI approach.

The PI approach is based on the well-known empirical Bayes (EB) principle [5–11]. (EB is a Bayesian approach, with the ‘prior information’ not to be supplied by the researchers, but estimated from the data itself.) Although the EB method has been developed for high-throughput data, most efforts are concentrated on microarray studies [12–18]; relatively few focus on (genetic) epidemiological studies [19–21]. In this paper, we extend the current EB methodologies in a number of ways. We allow for situations when the distribution of logORs is non-normal or unspecified, and then suggest the minimal number of factors (e.g., SNPs) to be analyzed together in a panel in order to guarantee the nominal coverage probability of a PI. To account for the phenomenon of linkage disequilibrium in genetic epidemiological studies, we also consider situations when the estimates of logORs are correlated with one another. The PI approach can enjoy straightforward probabilistic interpretations, and can achieve higher efficiency while maintaining nominal coverage probability.

Methods

Suppose that the logistic regression model in a study involves a total of m logORs (θ_i , $i = 1, \dots, m$). From the computer outputs, we obtain the point estimates (the regression coefficients), $\hat{\theta}_i$, and their standard errors, SE_i , for $i = 1, \dots, m$. If desired, the variance-covariance matrix of the regression coefficients can also be produced, with the diagonal elements of the matrix being the variances, $v_i = SE_i^2$ for $i = 1, \dots, m$, and the off-diagonal elements being the covariances, v_{ij} for $i \neq j$. For the i th logOR, the asymptotic 95% CI is $\hat{\theta}_i \pm 1.96 \times SE_i$. Note that the formulae for the CI and the following PI both require the asymptotic normality of MLEs. This may not hold true in studies with small sample sizes [22].

If these m factors are ‘exchangeable’, the following PI approach can be employed. The condition of exchangeability means that no extra information, other than the data itself, is available to distinguish some θ_i ’s from the others. For the i th logOR, the 95% PI is (see our Information S1 for the derivation):

$$\hat{\theta}_i^* \pm 1.96 \times \sqrt{\hat{v}_i^*},$$

where $\hat{\theta}_i^* = \left(\frac{\hat{\theta}_0}{\hat{v}_0} + \frac{\hat{\theta}_i}{v_i} \right) \left(\frac{1}{\hat{v}_0} + \frac{1}{v_i} \right)^{-1}$ (the posterior mean), $\hat{v}_i^* = \left(\frac{1}{\hat{v}_0} + \frac{1}{v_i} \right)^{-1}$ (the posterior variance), $\hat{\theta}_0 = \frac{1}{m} \sum_{i=1}^m \hat{\theta}_i$ (the prior mean), and $\hat{v}_0 = \max \left[\frac{1}{m-1} \sum_{i=1}^m (\hat{\theta}_i - \hat{\theta}_0)^2 - \frac{1}{m} \sum_{i=1}^m v_i + \frac{1}{m(m-1)} \sum_{i \neq j} v_{ij}, 0 \right]$ (the prior variance). From the formulae, we see that the individual

posterior mean is a weighted average of the prior mean and the individual MLE, and that the prior mean is the average of all individual MLEs ($\hat{\theta}_i$ ’s). Compared to the individual MLEs, the individual posterior means are therefore shrunk toward the overall mean. As for the variances, we see that the posterior variance is a harmonic mean of the prior variance ($\hat{v}_0$) and the individual variance (v_i). The posterior variance is thus guaranteed to be smaller than (with finite $\hat{v}_0$) or equal to (with infinite $\hat{v}_0$) the individual variance. This suggests an efficiency gain through ‘borrowing strength’ [23] from one another in that panel of the m exchangeable factors.

Before turning to the ‘adjusted’ (or ‘model-based’) ORs, the researchers may often examine the ‘crude’ ORs one by one for each and every factor in the study (some refer to this as ‘univariate analysis’). The above PI formula can also be used for these crude ORs, except that now the covariances are not directly from a model output, but can be approximated using the formula: $v_{ij} = r_{ij} \times SE_i \times SE_j$, where the r_{ij} is the Pearson’s correlation coefficient between the i th and the j th factors, and the SE_i and SE_j , the standard errors of the i th and the j th crude logORs, respectively.

As the name suggests, a PI is a prediction of the whereabouts of the true value. A 95% PI is a probabilistic prediction implying that with a probability of 0.95, the true value is within the region it delimits. This probability statement can be directly applied to the very sample the researchers are working on. The next section will examine these properties in detail.

Simulation Setup

We evaluate the performances of the PI approach in terms of average coverage probabilities. The average coverage probability for a total of m PIs is defined as:

Average Coverage Probability

$$\begin{aligned} &= \frac{\text{number of PIs that encompass the corresponding true values}}{\text{number of PIs in a study}} \\ &= \frac{1}{m} \cdot \sum_{i=1}^m I \left[\theta_i \in \left(\hat{\theta}_i^* \pm 1.96 \times \sqrt{\hat{v}_i^*} \right) \right], \end{aligned}$$

where $I[\text{statement}]$, an indicator function, has a value of 1 if the statement is true, and a value of 0 if otherwise. The average coverage probability is also the expected proportion of PIs encompassing their corresponding true values. We study the situation when the true values of the logORs (θ_i ’s) are normally distributed with a mean of 0 and a variance of v_0 . The v_0 is studied for values of 1 (95% of the ORs are between 0.14 and 7.10), 0.5 (between 0.25 and 4.00) and 0.25 (between 0.38 and 2.66), respectively. We also study the situations of non-normal distributions. To do this, we deliberately set the distribution of θ_i ’s to be very far from the normal: 40% of them coming from a beta distribution, 20% of them from a uniform distribution and the remaining 40% from a normal distribution. This mixture distribution is shown in our Information S1.

We also study situations when the estimates of the logORs ($\hat{\theta}_i$ ’s) themselves are independent of one another, as well as when they are correlated. We assume two levels for the correlations: moderate correlation (with a third of the pair-wise correlations coming from a uniform $[-0.4, 0]$ distribution, and the remainder from a uniform $[0, 0.4]$ distribution), and strong correlation (a third of the pair-wise correlations from a uniform $[-0.8, 0]$ distribution, and the remainder from a uniform $[0, 0.8]$ distribution). The variances of the estimates of the logORs (the

v_i 's) are studied for average values of $\bar{v}=0.5$ (uniform [0.25, 0.75] distribution; the ratio of upper limit to lower limit of 95% CI varies from $\exp[2 \times 1.96 \times \sqrt{0.25}] = 7.1$ to $\exp[2 \times 1.96 \times \sqrt{0.75}] = 29.8$), 0.25 (4.0 to 11.0) and 0.125 (2.7 to 5.5), respectively. $\bar{v}$ is related to the sample size (i.e., the number of subjects) of a study. Given a panel of factors, each with an exposure prevalence among controls drawn from a uniform [0.2, 0.8] distribution and $v_0=1$, a $\bar{v}$ of 0.5, 0.25 or 0.125 approximately corresponds to a sample size of 50, 100 and 200, respectively. The simulation is performed 10,000 times for each scenario investigated.

Results

Simulation Results

Figure 1 plots the (average) coverage probability against m , the number of logORs being considered for the PI approach, when θ_i 's are normally distributed (solid lines: independency; broken lines: moderate correlation; dotted lines: strong correlation). The

different panels in the figure are arranged such that the v_0 is 1, 0.5 and 0.25 (from top to bottom), and the $\bar{v}$ is 0.5, 0.25 and 0.125 (from left to right). We see that as more logORs are considered, their average coverage probability using the PI approach becomes closer to its nominal level of 95%. When $\bar{v}/v_0=0.125$ (panel C), m in the range of 10~20 suffices to achieve the desirable coverage properties for the PI approach. With $\bar{v}/v_0=0.25$ (panels B and F), $m \approx 30$ suffices. With $\bar{v}/v_0=0.5$, no more than 50 logORs are needed (panels A, E and I). For situations when $\bar{v}/v_0 > 0.5$, we need $m > 100$ to guarantee the desired coverage probability (panels D, H and G). Figure 2 shows the coverage probabilities of the PI approach when the distribution of θ_i 's is non-normal. We see that the results are very similar to those presented in Fig. 1.

As more attention in a study is often given to those ORs that are significant. (Some will refer to these as the 'discoveries' or 'findings' of the study), we also examine the coverage probabilities for those intervals that do not include zero (OR=1). Figure 3 shows the results for the PI approach (solid lines: independency; broken lines: moderate correlation; dotted lines: strong correlation).

Figure 1. Coverage probabilities when different numbers of logORs (m) are considered for the prediction interval approach, assuming logORs are normally distributed. The different panels in the figure are arranged such that the v_0 is 1, 0.5, and 0.25 (from top to bottom) and the $\bar{v}$ is 0.5, 0.25, and 0.125 (from left to right). Solid lines: independency among the estimates of the logORs; broken lines: moderate correlation among the estimates of the logORs; dotted lines: strong correlation among the estimates of the logORs.
doi:10.1371/journal.pone.0032022.g001

tion), and for the CI approach (constant solid lines), when θ_i 's are normally distributed. We see that even as we focus exclusively on those intervals that are 'significant', the intervals picked out by the PI approach can still guarantee the nominal coverage probability, that is, as the number of logORs is sufficiently large. Note that here we interpret the nominal coverage at its face value (95%, in our case) without resorting to multiple-testing adjustment of any kind. However, we see that the CI approach does not have this merit. In fact, significant intervals detected by the CI approach (without multiple-testing adjustment) can yield gross under-coverage (as low as 70%, in Fig. 3).

Application to Real Data

We apply our PI approach to a genetic epidemiological study on age-related macular degeneration. Klein et al. [24] reported a data set containing 96 cases and 50 controls. We chose the SNPs within 1 Mb from the complement factor H gene on chromosome 1q32, which was reported as a promising region for age-related macular degeneration [25]. In this region (192,917,342 bp —

195,266,213 bp), 66 SNPs are informative (minor allele frequency $\geq 1\%$), and conform to the Hardy-Weinberg equilibrium (with the Hardy-Weinberg exact p values ≥ 0.05 in the 50 controls).

A simple logistic regression model is fitted, one at a time, for each and every one of the 66 SNPs. At each SNP, the code is based on the number of copies of an arbitrarily chosen allele (0, 1 or 2). From the computer outputs, we obtain the point estimates (the regression coefficients), $\hat{\theta}_i$, and their standard errors, SE_i , for the i th logOR, for $i=1, \dots, 66$. We also calculate the r_{ij} , the Pearson's correlation coefficient between the i th and the j th SNPs. (The Pearson's correlation coefficients between allele counts (0, 1 or 2) at pairs of SNPs can measure the linkage disequilibrium between them [26–29].) Of the total 2145 pairs of SNPs, there are 1900 pairs with the absolute values of Pearson's correlation coefficients less than 0.2; 164 pairs between 0.2 and 0.4; 63 pairs between 0.4 and 0.8; and the remaining 18 pairs larger than 0.8. Using the formula presented in the Method section, we find $\hat{\theta}_0 = 0.050$, $\hat{v}_0 = 0.338$, $\bar{v} = 0.173$ and $\bar{v}/\hat{v}_0 = 0.512$. This example

Figure 2. Coverage probabilities when different numbers of logORs (m) are considered for the prediction interval approach, assuming the distribution of logORs is non-normal. The different panels in the figure are arranged such that the v_0 is 1, 0.5, and 0.25 (from top to bottom) and the $\bar{v}$ is 0.5, 0.25, and 0.125 (from left to right). Solid lines: independence among the estimates of the logORs; broken lines: moderate correlation among the estimates of the logORs; dotted lines: strong correlation among the estimates of the logORs.
doi:10.1371/journal.pone.0032022.g002

with a total of 66 SNPs is, therefore, amenable to the PI approach, since no more than 50 logORs are needed to guarantee the nominal coverage probability with $\bar{v}/\hat{v}_0 \approx 0.5$ (see panels A, E and I in Figs. 1 and 2). If attention is to be restricted to those intervals that are significant, normality assumption will have to be invoked. Because θ_i 's are unknown, we perform the Shapiro-Wilk normality test on $\hat{\theta}_i$'s in this example, and the p-value is 0.229.

Figure 4 shows the results using the CI approach (panel A), and the PI approach (panel B). The average length of the 95% CIs (for the logORs of the 66 SNPs) is 1.461, while the average length of the 95% PIs is 1.163, 20% shorter. Thus we see that the PI approach is statistically more efficient than the CI approach. Besides this, the PI approach has a straightforward probabilistic interpretation: within the regions delimited by the 66 PIs presented in the figure, one can expect to find $66 \times 95\% = 62.7$ true logOR values. One can also focus on the PIs for the significant SNPs (a total of 10 SNPs, marked with stars in Fig. 4). Within the regions delimited by these 10 PIs, one can expect to find $10 \times 95\% = 9.5$ true logOR values. Since the regions do not

contain any zeroes (true negatives), it means that the expected number of true positives among these 10 significant SNPs is 9.5.

Discussion

The PI approach is based on the EB principle with a long history [5–11]. However, traditional CIs still dominate epidemiological studies. Here we advocate using PIs to bound the uncertainty of association measures for a panel of exchangeable factors. (The 66 SNPs in the previous example are indistinguishable from each other a priori.) The exchangeability condition allows us to make the assumption that the true logOR values (θ_i 's) arise from a certain unknown/unspecified distribution, the key assumption of the PI approach. In this paper, we extend the current EB methodologies in the literature [23,30–32] in a number of ways. These include: 1) we present a simple closed-form formula for the PI approach; therefore there is no need to perform computer iterations and/or complex modeling; 2) we allow for the situations when the distribution of θ_i 's is non-normal or

Figure 3. Coverage probabilities for those intervals that do not include zero (OR = 1), for the prediction interval approach (solid lines: independency; broken lines: moderate correlation; dotted lines: strong correlation) and for the confidence interval approach (constant solid lines), when logORs are normally distributed.

doi:10.1371/journal.pone.0032022.g003

unspecified, and then evaluate the required number of factors to guarantee the probabilistic interpretations for PIs (Parametric assumptions of θ_i 's have been commonly made in the literature, for example, normal [23], mixture normal [32] or 'bell-shaped' [30,31].); and 3) we allow for situations when the estimates of θ_i 's are correlated with one another.

We need to emphasize that the PI approach is applicable to m exchangeable factors, where m is a required number of factors (e.g., SNPs) to guarantee the probabilistic interpretations for PIs. The required m is 10~20 when $\bar{v}/v_0=0.125$, 30 when $\bar{v}/v_0=0.25$, 50 when $\bar{v}/v_0=0.5$ and 100+ when $\bar{v}/v_0>0.5$. An applicable situation is to evaluate the associations of a panel of SNPs that are indistinguishable from one another a priori, as illustrated in our application. Other examples are panels of food items and occupational/environmental surveillance data, etc, where no extra information (other than the data itself) is available to distinguish some factors from others, and where the total number of factors investigated in a study is large enough for the PI approach. For example, a recent occupational surveillance on the New Zealand population was conducted to identify occupations that may contribute to the risk of non-Hodgkin's lymphoma [33]. Before the data analyses were conducted, a list of a priori high-risk occupations, including farmers, meat workers, painters, etc, was constructed based on previous literature. The EB method was then applied to the remaining hundreds of occupations having no a priori belief about their risks [33]. Also, we wish to point out that although we use the PI approach to bound the uncertainties of odds ratios in this paper, the methodologies can also be applied to

bound the uncertainties of other association measures, such as risk ratios and risk difference, etc.

When the PI approach is applicable to m exchangeable factors (m is large enough, as suggested above), it can provide not only a straightforward probabilistic interpretation, but also a more efficient interval inference than the CI approach. By 'more efficient', we mean smaller variances, and therefore shorter intervals while achieving the same probability to encompass the true value (though the confidence level of a CI relies on the 'repeated-sampling' interpretation). This efficiency gain comes from 'borrowing strength' [23] from one another in a panel of exchangeable factors. For the i th θ_i , the efficiency gain of the PI approach to the CI approach is

$$G_i = \frac{1/(\text{Variance of the PI approach}) - 1/(\text{Variance of the CI approach})}{1/(\text{Variance of the CI approach})}$$

$$= \frac{1/\hat{v}_i^* - 1/v_i}{1/v_i} = \frac{v_i}{\hat{v}_0}.$$

Because $v_i/\hat{v}_0 > 0$, theoretically, the PI approach will be more efficient than the CI approach. For a study with a very large sample size (v_i 's $\rightarrow 0$), the efficiency gain by using the PI approach may be marginal (G_i 's $\rightarrow 0$). However for a moderate-sized study, the efficiency gain can be considerable. In Figs. 1 and 2, the average efficiency gains are, diagonally from left lower to right upper panels, 200%, 100%, 50%, 25% and 12.5%, respectively. In the example of the age-related macular degeneration data (Fig. 4), the average efficiency gain is

Figure 4. Results using the confidence interval approach and the prediction interval approach. (A) The 95% confidence intervals (CIs) and (B) the 95% prediction intervals (PIs) of the natural logarithms of odds ratios for the 66 single nucleotide polymorphisms (SNPs) in the age-related macular degeneration data. In (B), stars identify the 10 significant SNPs with corresponding 95% PIs not including zero.
doi:10.1371/journal.pone.0032022.g004

51.2%. In this paper, we focus on ‘interval estimates’ and show that the interval estimates of the EB-based PI approach have desirable statistical properties.

It is worth noting that the EB-based ‘point estimates’ are a different story: they have lower mean squared errors, but are biased by themselves [34]. Although the point estimates are not statistically unbiased, the EB-based adjustment can weed out false positives by shrinking imprecise outliers toward the overall mean. Through this shrinkage process, true positives may remain unambiguously positive, while false positives are likely to be removed (as shown by a large surveillance data set of occupation and cancer [35]). Using the EB-based adjustment in genetic studies is justifiable because follow-up or replication driven by false-positive findings often leads to a considerable cost.

The PI approach can provide a valid statistical inference for multiple tests. Treating multiple PIs as multiple tests, we show in the simulation that for those significant intervals identified by the PI approach (Fig. 3), the nominal coverage can be retained at its face value (95%, in our case) when θ_i ’s are normally distributed. This suggests a link between the EB-based PI approach and the multiple-testing control of false discovery rate (FDR, defined as the expected proportion of false rejections among those that reject the null hypothesis of no association [36]). In fact, this connection follows directly from the Bayes theorem [13]. We apply Storey and Tibshirani’s algorithm of FDR control [37] to the age-related

macular degeneration data, and also identify the same 10 SNPs as detected by the PI approach. (The expected number of true positives among these 10 SNPs is 9.85, using Xie et al.’s permutation method [38].) More work needs to be done for multiple-testing controlling properties of the PI approach when the distribution of θ_i ’s is non-normal.

Supporting Information

Information S1 The derivation of the prediction interval approach, the distribution of θ_i and e^{θ_i} for the non-normal situation, and Figure 3 with y-axis ranging from 0.8 to 1.0.

(DOC)

Acknowledgments

We thank the anonymous reviewers for their constructive comments. We also thank Dr. Josephine Hoh for kindly providing the data set of age-related macular degeneration.

Author Contributions

Conceived and designed the experiments: WCL. Performed the experiments: WYL. Analyzed the data: WYL. Contributed reagents/materials/analysis tools: WYL. Wrote the paper: WYL WCL.

References

- Kraemer HC (2004) Reconsidering the odds ratio as a measure of 2×2 association in a population. *Stat Med* 23: 257–270.
- Rothman KJ, Greenland S, Lash TL (2008) *Modern epidemiology*. 3rd Ed Lippincott Williams & Wilkins.
- Agresti A (2003) Dealing with discreteness: making ‘exact’ confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Stat Methods Med Res* 12: 3–21.
- Cox DR, Hinkley DV (1974) *Theoretical Statistics* Chapman & Hall.
- He K (1992) Parametric empirical Bayes confidence intervals based on James-Stein estimator. *Statistics & Decisions* 10: 121–132.
- Laird NM, Louis TA (1987) Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J Am Statist Ass* 82: 739–757.
- Morris C (1983) Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis and Robustness*. New York: Academic Press. pp 25–50.
- Morris C (1983) Parametric empirical Bayes inference: theory and applications. *J Am Statist Ass* 78: 47–55.
- Robbins H (1956) An empirical Bayes approach to statistics. *Proc Third Berkeley Symp on Math Statist and Prob (Univ of Calif Press)* 1: 157–163.
- Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis* 2nd Ed Chapman & Hall/Crc.
- Laird NM, Louis TA (1989) Empirical Bayes confidence intervals for a series of related experiments. *Biometrics* 45: 481–495.
- Efron B (2009) Empirical Bayes Estimates for Large-Scale Prediction Problems. *J Am Stat Assoc* 104: 1015–1028.
- Efron B, Tibshirani R (2002) Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 23: 70–86.
- Huang T, Tu K, Shyr Y, Wei CC, Xie L, et al. (2008) The prediction of interferon treatment effects based on time series microarray gene expression profiles. *J Transl Med* 6: 44.
- Hwang JT, Liu P (2010) Optimal tests shrinking both means and variances applicable to microarray data analysis. *Stat Appl Genet Mol Biol* 9: Article36.
- Datta S, Datta S (2005) Empirical Bayes screening of many p-values with applications to microarray studies. *Bioinformatics* 21: 1987–1994.
- Datta S, Satten GA, Benos DJ, Xia J, Heslin MJ, et al. (2004) An empirical bayes adjustment to increase the sensitivity of detecting differentially expressed genes in microarray experiments. *Bioinformatics* 20: 235–242.
- Lin Y, Reynolds P, Feingold E (2003) An empirical bayesian method for differential expression studies using one-channel microarray data. *Stat Appl Genet Mol Biol* 2: Article8.
- Chen GK, Witte JS (2007) Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet* 81: 397–404.
- Heron EA, O’Dushlaine C, Segurado R, Gallagher L, Gill M (2011) Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. *Biostatistics* 12: 445–461.
- Stromberg U (2009) Empirical Bayes and semi-Bayes adjustments for a vast number of estimations. *Eur J Epidemiol* 24: 737–741.
- King EN, Ryan TP (2002) A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician* 56: 163–170.
- Hwang JT, Qiu J, Zhao Z (2009) Empirical Bayes confidence intervals shrinking both means and variances. *J R Statist Soc B* 71: 265–285.
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308: 385–389.
- Lin W-Y, Lee W-C (2010) Incorporating prior knowledge to facilitate discoveries in a genome-wide association study on age-related macular degeneration. *BMC Research Notes* 3: 26.
- Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, et al. (2005) Characterization of multilocus linkage disequilibrium. *Genet Epidemiol* 28: 193–206.
- Schaid DJ (2004) Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 166: 505–512.
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35: 235–254.
- Zaykin DV (2004) Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet Epidemiol* 27: 252–257.
- Greenland S (1993) Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med* 12: 717–736.
- Greenland S, Poole C (1994) Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. *Arch Environ Health* 49: 9–16.
- Qiu J, Hwang JT (2007) Sharp simultaneous confidence intervals for the means of selected populations with application to microarray data analysis. *Biometrics* 63: 767–776.
- Mannetje A, Dryson E, Walls C, McLean D, McKenzie F, et al. (2008) High risk occupations for non-Hodgkin’s lymphoma in New Zealand: case-control study. *Occup Environ Med* 65: 354–363.
- Zhao Y, Lee AH, Barnes T (2010) On application of the empirical Bayes shrinkage in epidemiological settings. *Int J Environ Res Public Health* 7: 380–394.
- Steenland K, Bray I, Greenland S, Boffetta P (2000) Empirical Bayes adjustments for multiple results in hypothesis-generating or surveillance studies. *Cancer Epidemiol Biomarkers Prev* 9: 895–903.
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57: 289–300.
- Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100: 9440–9445.
- Xie Y, Pan W, Khodursky AB (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 21: 4280–4288.