Bayesian analysis of surveillance data

Kyle Steenland

What's really significant?

Empirical and semi-Bayes
adjustments in data with multiple outcomes

Kyle Steenland
Emory School of Public Health
Emory University, Atlanta
email nsteenl@sph.emory.edu

14, September 2002, EPICOH2002, Barcelona

Traditional methods of adjustment for multiple comparisons (eg., the Bonferroni adjustment) have fallen into disuse. It has been argued that such multiple comparison adjustments are unnecessary and in fact ill-advised, because they assume a global null hypothesis which is neither plausible nor of interest, and because they are too conservative and may lead investigators to ignore unexpected but important findings (Rothman, 1990).

When faced with a large number of comparisons, many epidemiologists currently do no adjustment at all, but instead use whatever a priori knowledge exists, as well as common sense and biological plausibility, to evaluate what findings are important in their data.

However, Greenland and Robins (1991) and Greenland and Poole (1994) have argued that in some circumstances empirical or semi-Bayes adjustments can be useful as an alternative to traditional multiple comparison adjustments. These circumstances are that 1) a large number of comparisons are made, 2) the comparisons can be grouped into sets within which all comparisons can be considered similar or "exchangeable", 3) random error is present and presumably accounts for much of the observed variation in the parameters estimated to evaluate the comparisons (eg., relative risks, rate ratios, regression coefficients), and 4) investigators must choose which comparisons to investigate further, and there is a significant cost to such further investigations.

Possible applications:

Surveillance data: eg, death certificates and occupation

Large scale gene-environment interaction results: multiple outcomes, exposures, and genetic polymorphisms

A single study with a large number of multiple outcomes, eg., a case-control study of colon cancer and diet in which one considers 100 dietary items as exposures

The basic idea of empirical Bayes (EB) adjustments for multiple associations, such as log relative risks, is that the observed spread or variation of the estimated relative risks around their mean is larger than variation of the true but unknown log relative risks.

Empirical Bayes adjustments attempt to estimate this extra variation from the data at hand, and then use this estimate to adjust the observed relative risks. Typically, this adjustment serves to pull or shrink outlying log relative risks in towards their mean, more so if the estimate to be adjusted has a large individual variance.

This shrinkage attempts to anticipate "regression to the mean," in which outlier observations tend to shrink toward (ie, become closer to) the mean upon obtaining new data. A consequence of this shrinkage is that the overall variance of the EB-adjusted estimates is smaller than that of the unadjusted estimates. The variance of each estimated relative risk is also re-estimated.

Although the individual EB-adjusted estimates are not statistically unbiased, the average squared error of the adjusted estimates will generally be less than the average squared error of the original estimates (trade off: bias vs precision). Empirical Bayes estimators are part of a class of "shrinkage" estimators with a long history in the statistical literature

In semi-Bayes adjustments the investigator specifies or chooses a reasonable a priori value for the extra variation, instead of estimating it from the data.

Empirical and semi-Bayes adjustments can be thought of as constraints on empirical epidemiologic findings. These constraints serve to reduce the observed variability of multiple outcomes. The constraints can be based solely on the observed data (empirical Bayes) or partly on a priori considerations (semi-Bayes).

New point estimates and confidence intervals are generated after using these constraints.

General considerations: our example is based on hierarchical models for multiple RRs when these RRs are thought to be "exchangeable", all stemming from some common source population

model 1: $\log RR_i \sim N(1_i, \Phi_i^2)$

model 2: $1_i \sim N(:, 9^2)$

where model 2 is the assumed prior distribution of the true but unknown (true) 1_is

We might assume a prior distribution of

:=0 and we might estimate 9^2 from the observed data (empirical Bayes) or some prior knowledge of the likely variation in the observed RRs (semi-Bayes)

General considerations: Bayes rule and Bayesian inference

Bayes rule: $Pr(y/x) = pr(x/y)pr(y)/pr(x)$

or if $y_i = 1_i$=log relative risk$_i$, x=the observed data, and pr(1) is some assumed prior distribution of 1, then rewriting we have

$Pr(1/x) = pr(x/1)pr(1)/pr(x)$,

the pr(x) is taken as a constant, and pr(x/1) is the usual likelihood of the data given 1, then the posterior distribution of 1 given the data is given

$Pr(1/x) \sim$ Likelihood(x/1)pr(1)

So the posterior distribution of 1/x is a function of the usual (frequentist) likelihood and the prior (Bayesian) distribution of 1

Estimate of any given posterior 1_i will be a weighted average of the MLE for 1_i and the mean of the prior distribution.

Specific considerations

Assume we have a large number of "exchangeable" log RR$_i$'s (we have no a priori reason to believe any specific log RR$_i$ differs from 0). Assume these RR$_i$'s are centered around the null (mean log RR$_i$=0), but have some observed spread or variation. Let that observed variation be Var_{obs}

Assume that the unknown 'true' log RRs (1_is) have some observed spread or variation called Var_{true} (ϑ^2)

We can assume that $Var_{obs} > Var_{true}$, because of random error in measuring the true.

Let the average variance of the individual log RR_i's be Var_{mean}.

Then one can derive the following approximation:

$Var_{true} \cong Var_{obs} - Var_{mean}$

or $Var_{obs} \cong Var_{true} + Var_{mean}$

This implies that the observed variance among the log RRs is the sum of random error of each log RR_i and the true variance of log RRs.

Also, for procedure to work, $Var_{obs} - Var_{mean}$ must be >0, since $Var_{true} > 0$.

Let $lnRR_{mean} = \Sigma (lnRR_i * w_i) / \Sigma w_i$

The weights have the form

$$w_i = 1 / (S_i^2 + Vhat_{true}),$$

where S_i^2 is the variance of each $lnRR_i$, and $Vhat_{true}$ is the estimated Var_{true}. Note that since $Vhat_{true}$ is the result of these calculations in an EB analysis, we cannot know it at the beginning (Var_{true} is specified at the start of a semi-Bayes analysis). Thus empirical Bayes analyses require the use of iteration, where an initial guess of Var_{true} is used and then this initial guess is refined iteratively.

Now let

$$D = lnRR_i - lnRR_{mean}$$

and

$$Vhat_{obs} = \Sigma (w_i * D_i^2) / \Sigma w_i .$$

$Vhat_{obs}$ is our estimate of Var_{obs}. Then derive the timate for Var_{mean} as follows:

$at_{mean} = \Sigma (w_i * S_i^2) / \Sigma w_i .$

can now estimate Var_{true} by

$at_{true} = Vhat_{obs} - Vhat_{mean}$,

y max $(Vhat_{obs} - Vhat_{mean}, \tau^2)$ where τ^2 is a -specified minimum plausible value for Var_{true} i-Bayes method).

Finally, we can derive our empirical Bayes estimate of each true RR_i as a weighted average of the original estimate and the mean of the estimates as follows:

$$EB_i = ((Vhat_{true} * \ln RR_i) + (S_i^2 * \ln RR_{mean})) / (Vhat_{true} + S_i^2).$$

Note that if $Vhat_{true}$ is large this gives more weight to the original estimate. On the other hand, if the variance of the individual estimate S_i^2 is large, this gives more weight to the overall mean of the estimates.

Outlier log RR's will be pulled in or regressed to their mean; the degree of pulling in will depend on both the individual estimated variance of the outlier (more variance, more pulling in), and the estimated true variance ($Vhat_{true}$). The final or shifted new distribution log EB_is will have a smaller spread or variance than the original RR_is.

Example:

Nordic cancer/occupation study

Based on 1970 Census, ages 20-64, 10 million people

Norway, Sweden, Denmark, Finland

Follow-up through 1990 via cancer registry

Calculation of SIRs (standardized incidence ratios) for 34 sites of cancer, 54 occupational groups, men and women separately

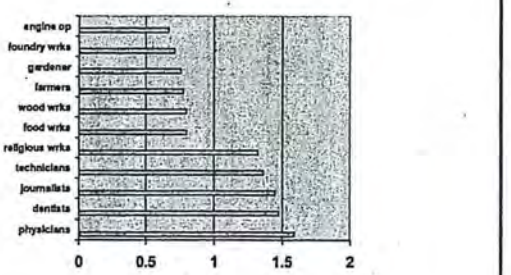
Problem of confounding of occupational associations by non-occupational risk factors linked to social class

Eg, women in high SES occupations have more breast cancer due to different reproductive patterns, men in low SES occupations have more lung cancer due to smoking habits.

Too offset this problem, we restricted data to manual laborers/crafstmen, which reduced SES confounding.

This group would be expected to have the most exposure to occupational carcinogens

Figure 1. Highest and lowest SIRs for female breast cancer



Re-scaling

Among manual laborers/craftsmen, we re-scaled the SIRs for each cancer so that they would overall be set to 1.00 across all occupations in the group.

Eg, if the overall lung cancer SIR for manual laborers was 1.5, we multiplied all lung cancer SIRs in for manual laborers by 1/1.5, so the overall lung cancer SMR adjusted would be centered on 1.0. Variance of the log SIR not changed by this adjustment.

Table 1. Summary of cancer and occupation standardized incidence ratios (SIRs) for manual laborers/craftsmen with $p<0.05$, adjusted for social class, before and after EB adjustment

	Positive*, before EB Adjustment	Negative*, before EB adjustment	Positive, after EB adjustment	Negative, after EB adjustment
Total, 1015	107 (11%)	68 (7%)	71 (7%)	37 (4%)
Men, 642	84 (13%)	54 (8%)	62 (10%)	33 (5%)
Women, 373	23 (6%)	14 (4%)	9 (2%)	4 (1%)

* Positive means SIR>1.0, negative means SIR<=1.0

Figure 2a. Original P-values for SIRs for manual laborers/craftsmen, males

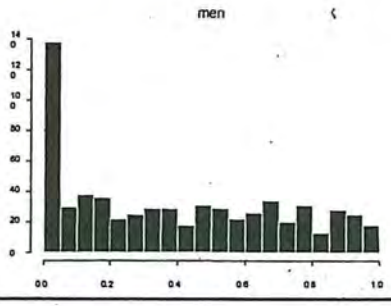


Figure 2b. P-values for SIRs for males, after empirical-Bayes adjustment

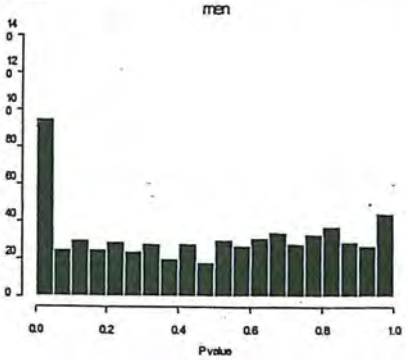


Figure 3a. Original P-values for SIRs for manual laborers/craftsmen, females

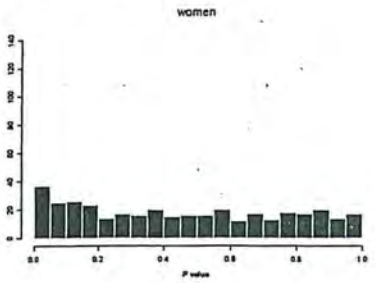


Figure 3b. P-values for SIRs for females, after empirical-Bayes adjustment

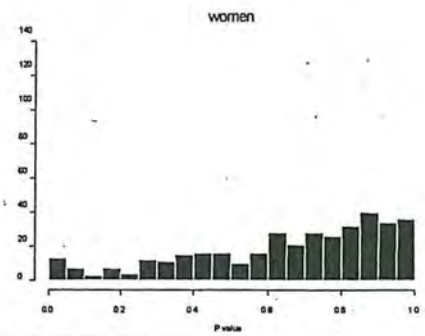


Figure 3c. P-values for SIRs for females, after semi-Bayes adjustment (assuming true variance of lnSIRs for females equal to that of males, ie., 0.019)

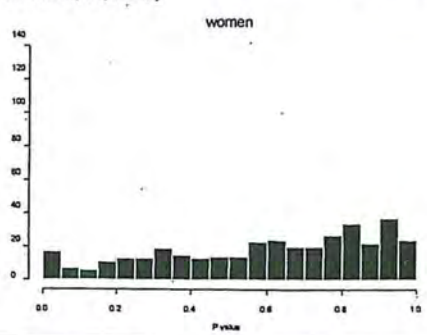


Figure 4a. Manual laborers/craftsmen males: lnSIRs, before empirical Bayes adjustment

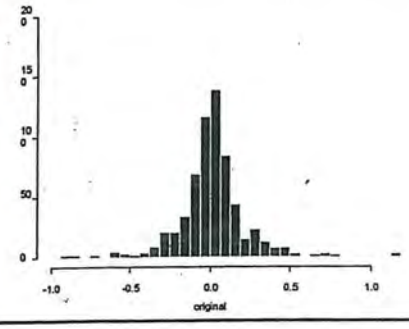


Figure 4b. Manual laborers/craftsmen males: lnSIRs, after empirical Bayes adjustment

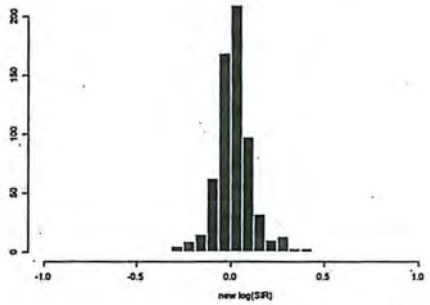


Figure 5a. Manual laborers/craftsmen females: lnSIRs, before empirical Bayes adjustment

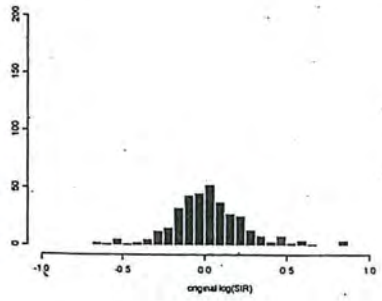


Figure 5b. Manual laborers/craftsmen females: lnSIRs, after empirical Bayes adjustment

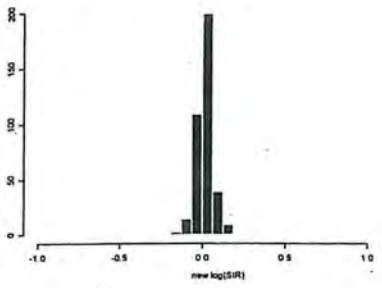


Figure 5c. Manual laborers/craftsmen, females: lnSIRs after semi-Bayes adjustments (assuming true variance of lnSIRs for females equal to that of males, ie., 0.019)

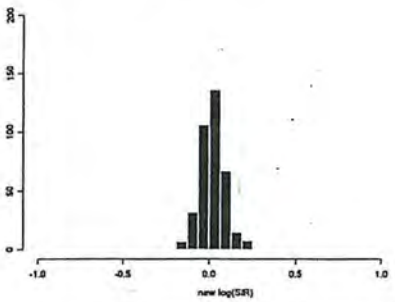


Table 2. Selected positive (SIR>1.0) findings suspected a priori, supported by empirical Bayes (EB) adjustment (males)

Occupation/cancer	Social class adjusted SIR (95 % CI)	Social class and EB adjusted SIR (95% CI)
Plumber/pleura	3.21 (2.52-4.09)	1.96 (1.62-2.36)
Woodworkers/nasal	1.62 (1.39-1.90)	1.46 (1.27-1.67)
Miners/lung	1.37 (1.27-1.48)	1.34 (1.24-1.45)

Table 3. Selected positive (SIR>1.0) findings not suspected a priori, not supported by empirical Bayes (EB) adjustment (males)

Occupation/cancer	Social class adjusted SIR (95% CI)	Social class and EB adjusted SIR (95% CI)
Chimney sweep/liver	2.17 (1.06-4.43)	1.12 (0.87-1.45)
Printer/breast	2.08 (1.10-3.93)	1.14 (0.88-1.47)
Beverage worker/oral	2.25 (1.10-4.60)	1.13 (0.87-1.46)

Table 4. Selected positive (SIR>1.0) findings not suspected a priori, supported by empirical Bayes adjustment

occupation/cancer	Sex	Social class adjusted SIR (95% CI)	Social class and EB adjusted SIR (95% CI)
Mechanic/breast	M	1.41 (1.09-1.83)	1.21 (1.01-1.47)
Construction/lip	M	1.48 (1.35-1.63)	1.42 (1.30-1.56)
Welder/NHL	M	1.23 (1.06-1.42)	1.17 (1.03-1.34)

Closing thoughts

Surveillance studies will group many people not exposed to an occupational carcinogen with those actually exposed, generally diluting any true risks. It would therefore be expected that exposure-related elevations of SIRs would be modest.

Also, most strong occupational carcinogens had been discovered and exposures lowered by the 1980s in the industrialized countries. So effects of new true carcinogens are likely to be modest.

This means that surveillance studies of occupation and cancer may be unlikely to discover new true risks.

Nonetheless, Bayesian techniques will likely improve over usual methods.

16th **EPICOH**
Congress on Epidemiology in Occupational Health

and

Jack Pepys Symposium on Prevention of Occupational Asthma

and

3rd International Congress on Women's Health: Occupation, Cancer and Reproduction

Course on New methods in Epidemiology

Saturday, 14 September 2002

Barcelona, Spain

2002