Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System

William DuMouchel

A common data mining task is the search for associations in large databases. Here we consider the search for "interestingly large" counts in a large frequency table, having millions of cells, most of which have an observed frequency of 0 or 1. We first construct a baseline or null hypothesis expected frequency for each cell, and then suggest and compare screening criteria for ranking the cell deviations of observed from expected count. A criterion based on the results of fitting an empirical Bayes model to the cell counts is recommended. An example compares these criteria for searching the FDA Spontaneous Reporting System database maintained by the Division of Pharmacovigilance and Epidemiology. In the example, each cell count is the number of reports combining one of 1,398 drugs with one of 952 adverse events (total of cell counts = 4.9 million), and the problem is to screen the drug-event combinations for possible further investigation.

KEY WORDS: Adverse drug reactions; Association; Gamma-Poisson model; Mixture model; Shrinkage estimate.

1. THE PROBLEM: FINDING "INTERESTINGLY LARGE" CELL COUNTS IN A LARGE FREQUENCY TABLE

Consider a large frequency table formed by considering combinations of categorical variables (attributes) from a database. Here "large table" means one having very many cells, either because there are many variables or because some of the variables have very many categories. This article focuses on the latter situation and considers the case where two of the variables each have 1,000 or so categories and the number of cells (combinations of variable values) is in the millions. The methodology discussed here is directly applicable to alternative situations, however, in which the millions of cells might be determined by combinations of more variables, each having fewer categories. To be more specific, suppose that A (having a levels, $A = 1, \ldots, a$) and B (having b levels, $B = 1, \ldots, b$) are two variables of most interest, and that $C(= 1, \ldots, c)$ is a third, stratification variable, of less direct interest, but that may be associated with A and/or B. That is, primary interest centers on the $M = ab$ combinations of A and B, but all abc counts N_{ijk}, the frequencies in the data base of $A = i$, $B = j$, $C = k (i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, c)$, are available.

This setup is relevant to many data mining applications. Hand (1998) discussed the relationship between data mining and statistics. Consider the following three problems from quite different application areas.

1. *Construction of a lexicon of phrases for use in natural language programming*. The goal is to find pairs of words that very often occur together in text databases. The application might be one of computer text or speech understanding, or alternatively the construction of efficient document search procedures within the World Wide Web. In this case (A, B) are (Word1, Word2) in tabulations of counts of consecutive words within a corpus of text, and N_{ijk} is the count of (Word1 $= i$, Word2 $= j$) in the kth body of text.

2. *Marketing analyses of supermarket scanner data*. The goal is to find pairs of products that are often purchased together by shoppers. In this case $N_{ijk} (i < j)$ could be the number of times any shopper bought both product i and product j at the kth store location.

3. *Screening reports of adverse drug reactions*. We shall focus on this example. As described in more detail in Section 4, the U. S. Food and Drug Administration (FDA) Spontaneous Reporting System (SRS) database consists of reports of medical events happening to patients taking various drugs. This article presents an analysis in which N_{ijk} are the counts of the number of reports involving drug i (having $a = 1398$ levels) and event j (having $b = 952$ levels) stratified by $c = 18$ combinations of report date and reported gender of the patient. (There were six five-year time period groupings and three gender report classifications: male, female, and unreported gender.) Thus, $M = 952 \times 1398 = 1{,}330{,}896$ cells (combinations of event and drug) are of direct interest, although the cell counts are separately available for each time period-reported gender stratum. The goal is to screen all the drug-event combinations for possible further investigation. The following de-

William DuMouchel is a Technology Consultant, AT&T Labs—Research, Florham Park, NJ (Email: dumouchel@research.att.com). The author thanks David Fram, Ana Szarfman, and Jonathon Levine for introducing him to the SRS database; Sally Cassells and Barbara Snow for extracting and preparing the required data summaries from it; Ilya Yunus for extensive programming assistance; and Colin Mallows and Robert O'Neill for comments on earlier drafts. The C program GPS was produced by Belmont Research, Inc., under a contract with the U.S. FDA Center for Drug Evaluation and Research, with funding provided by a grant from the Office of Women's Health, Ana Szarfman, principal investigator. The algorithms in GPS are based on S-Plus programs written by William DuMouchel with support from Columbia University and AT&T Labs.

scription is a slightly abridged version of material posted at the Internet site http://www.fda.gov/cder/adr/.

BACKGROUND: The Spontaneous Reporting System (SRS) of the Division of Pharmacovigilance and Epidemiology (DPE) is a computerized database of adverse drug reactions (ADRs) primarily reported by health professionals. The present database contains over 1 million reports collected since 1969. The system contains only adverse events detected and reported after marketing of the drug. The primary purpose for maintaining the database is to serve as an early warning or signaling system for adverse drug reactions not detected during premarketing testing.

The SRS depends on the detection of a new clinical event, the attribution of the clinical event to the administration of a drug, and the reporting of that event to a drug company or the FDA.

The health professional may choose to report the adverse reaction to a drug firm, who must, by law, report the information to the FDA. Ninety percent of reports are received from drug manufacturers. DPE receives the remaining ten percent directly from other reporters (i.e. health professionals and consumers).

Data from all reports are entered into the DPE adverse drug reaction database and the reports are scanned into an electronic filing system.

LIMITATIONS:
THERE ARE IMPORTANT THINGS TO REMEMBER WHEN REVIEWING OR ANALYZING DATA FROM THE SPONTANEOUS REPORTING SYSTEM.
1. For any given report, there is NO CERTAINTY that the suspected drug caused the reaction. Physicians are encouraged to report suspected reactions. The event may have been related to the underlying disease for which the drug was given, to concomitant drugs being taken, or may have occurred by chance at the same time the suspected drug was taken.
2. The number of reports is not equal to the number of people/patients. Multiple reports are often received for the same patient-event from the same or different sources.
3. Accumulated case reports cannot be used to calculate true incidence or estimates of drug risk. Substantial under-reporting to the FDA occurs.
4. Numbers from these data must be carefully interpreted as reporting rates and not occurrence rates. True incidence rates cannot be determined from these data. Comparisons of drugs should be made with caution because many factors may result in differential reporting of adverse events by drug.

1.1 Searching for Unusually Frequent Drug-Event Combinations

The limitations of the SRS data raise many questions. Can drug-event combinations of potential interest be identified from internal evidence alone? How can a rate be defined without a denominator? Although the Spontaneous Reporting System has served the valuable purpose of helping detect serious adverse drug reactions not detected during premarketing testing, the full potential value of this warning system has not been realized because of the difficulty of interpreting the reported frequencies. Unlike a well-designed clinical trial or epidemiology study, the SRS does not allow computation of incidence rates or dose-response curves for a given combination of drug and adverse event. The problem of computing such a rate turns upon finding an appropriate denominator with which to compare the frequency of a reported combination. The information infrastructure is not in place to allow reliance on wholesale or retail sales data or on national counts of prescriptions for each drug. Even if such aggregate data were available, the usual biostatistical analyses would require data on the distributions of age, sex, medical condition, and actual drug consumption patterns of the millions of drug consumers who do not report adverse events, as well as of those who do. In addition, the previously mentioned variations in reporting rates and reporting styles across drug manufacturers and adverse event definitions, which is only to be expected in a system lacking expensive formal research-style protocols, also make interpretation difficult. Rather than attempt to match up the reported frequencies (N_{ij} for drug i and event j) in the SRS with an external measure of exposure, the methodology of this article uses an internally derived measure, namely *baseline frequencies* E_{ij}. Using the baseline frequencies as a denominator, the relative report rates, $\mathrm{RR}_{ij} = N_{ij}/E_{ij}$, are the statistics of interest, since they measure how many times more frequently the combination of drug i and event j has been reported than would be expected to occur if reports involving drug i are statistically independent of reports involving event j. Of course, RR is not always a valid detection statistic. To take an extreme hypothetical, suppose a given drug i is so deleterious that taking it magnifies the probability of every conceivable adverse event by a factor of 10, compared to background incidences. Then RR_{ij} will still equal 1, not 10, for every j, since that drug will not be associated with any particular adverse events, and the result is the same as if the usage of an innocuous drug were 10 times as great. But usually drugs cause at most a few types of adverse events, compared to the wide variety of all reported events, so that such causality will lead to increases in the corresponding RR_{ij}.

Unfortunately, the converse inference is not so reliable, since there are many possible causes of differential reporting of adverse events by drug, so that a large value of RR_{ij} is only a possible indicator of a medical cause and effect relationship. The unreliability of RR comes from two quite different sources—unreliability due to sampling variance and unreliability due to reporting biases. The empirical Bayes methodology developed in the following is quite good at minimizing the effect of sampling variance on the interpretation of RR, but does nothing to minimize reporting bias. Therefore the method should be thought of as a way of screening drug adverse event rates occurring in the SRS, but combinations so identified still require the application of medical and epidemiological sophistication before public health pronouncements are appropriate. As discussed by Hand (1998), many data mining problems involve a mixture of statistical modeling and more ad-hoc application of expert judgement, even when the sample size is large.

The recommended methodology for evaluating the cell counts N_{ij} in data mining applications like the ones described consists of two steps: the definition of a baseline frequency E_{ij} and the definition of a measure for comparing the N_{ij} to the E_{ij}.

2. DEFINING THE BASELINE FREQUENCIES

The following notation is useful:

$$N_{ij} = N_{ij.} = \sum_k N_{ijk}$$

$$N_{i.k} = \sum_j N_{ijk}$$

$$N_{.jk} = \sum_i N_{ijk}$$
$$N_{..k} = \sum_i \sum_j N_{ijk}$$
$$E_{ij} = \sum_k N_{i.k} N_{.jk} / N_{..k}. \quad (1)$$

The quantity E_{ij} is denoted the *baseline* or null hypothesis frequency for cell $(A = i, B = j)$. It is the expected count assuming that the variables A and B are independent, conditional on C [see, for example, Agresti (1990, chap. 6)], and is also the expected count used by the Cochran (1954) and Mantel–Haenszel (1959) methods for combining multiple two-way tables in a test for independence. However, the purpose here is not to test for independence, since it is taken for granted that A and B are associated, but to establish a measure for comparing the M cell counts: N_{ij} is only "interestingly large" if it is large compared to E_{ij}. The definition of E_{ij} may change from that in (1), depending on the application. For example, in the marketing analysis of product purchases mentioned earlier, the fact that only sets rather than ordered pairs of products are of interest (i.e., only counts N_{ij} where $i < j$ are tabulated) will cause corresponding adjustments to the calculation of baseline frequencies. Clearly, the definition of E must depend on the dimensionality of the frequency table, the meaning of the variables being studied, and the purpose of the data mining project.

Three general principles relate to computational, statistical, and substantial considerations, respectively. Computationally, the calculations to compute E should scale up manageably if the methodology is to apply to very large databases. Often the fitted values from a graphical model (see, e.g., Almond 1995; Edwards 1995a; Lauritzen 1996; Pearl 1988) will be a sensible and computationally efficient choice for E. Statistically, the main consideration in the choice of how to define the baseline frequencies is that the variances of the E_{ij} should be much less than those of the N_{ij}, because the various methods for computing measures for comparing the Ns to the Es in step two of this screening methodology ignore the sampling variance involved in computing the Es and instead treat them as known constants. In the specific example of Equation (1), the (relative) variances of the E_{ij} are approximately inversely proportional to the sizes of the one-dimensional marginal totals $N_{i..}$ and $N_{.j.}$. If both a and b are large, then the variances of the E_{ij} will be much less than that of the N_{ij}.

As a substantial consideration, the baseline frequency definition should be easily interpretable so that deviations of the Ns from the Es are easily interpretable, and so that cells having N_{ij} approximately equal to E_{ij} are probably not "interesting," while greater increases of N_{ij} above E_{ij} make the cell "more interesting." In the case of Equation (1), the derivation of the baseline frequencies as arising from assuming the conditional independence of A and B given C, would usually satisfy this interpretation requirement. It is assumed that any association among categories of A and B is not of interest if it arises merely because of these variables' mutual association with C; that is, one wants to avoid being misled by the well-known Simpson's (1951) paradox. If this is not true, and associations due to mutual association with C are of equal interest, then a nonstratified definition $(E_{ij} = N_{i..} N_{.j.} / N_{...})$ may be preferable. Note that this article is neglecting the possibility that abnormally small counts, in which $N_{ij} \ll E_{ij}$, are of primary interest.

3. MEASURES COMPARING OBSERVED AND BASELINE COUNTS

Once a formula for baseline frequency is determined, step two of the methodology consists of selecting a scalar function of N and E and then ranking all the cells according to this function. That is, the choice of function defines a scale for being "interesting." Three conceptually different scales, labeled *relative risk*, *statistical significance*, and *empirical Bayes* are investigated and compared in this article.

3.1 Relative Risk or Relative Report Rate

Perhaps the simplest criterion is based on the ratio

$$\mathrm{RR}_{ij} = N_{ij} / E_{ij}. \quad (2)$$

The relative risk measure has the great advantage of being easy to interpret. No statistical or probabilistic calculations are involved in computing these measures, and the ratio measure is appealing: for example, if $N_{ij}/E_{ij} = 1{,}000$, then cell (i, j) occurred 1,000 times as frequently as the baseline frequency predicts. Its biggest disadvantage is the extreme sampling variability of RR when baseline and observed frequencies are small. The values $N = 1, E = .001$ have a very different statistical interpretation than $N = 100, E = .1$, even though both lead to $\mathrm{RR} = 1{,}000$. A statistic equivalent to RR, discussed by Church and Hanks (1991) is denoted by them as the mutual information statistic, $I_{ij} = \log_2(\mathrm{RR}_{ij})$. (Unfortunately, Dunning (1993) also used the term "mutual information" to denote a statistic more closely related to the significance test family discussed in the next subsection.)

Another measure similar in spirit to RR is the conditional probability measure used in Friedman et al. (1995) and discussed in DuMouchel et al. (1996), namely $\mathrm{CP}_{ij} = N_{ij} / \min(N_{i.}, N_{.j})$, which is an estimate of the maximum of $P(A = i | B = j)$ and $P(B = j | A = i)$.

3.2 Statistical Significance

Sampling variability can be explicitly addressed by using a statistical test criterion for testing the null hypothesis that $E[N_{ij}] = E_{ij}$. The measure LogP is defined as

$$\mathrm{LogP}_{ij} = -\log_{10}\left(\Pr\left[X \geq N_{ij}\right]\right), \quad \text{where} \quad X \sim \mathrm{Poisson}(E_{ij}). \quad (3)$$

Other measures can be derived by considering other test statistics. If E_{ij} is large, the normal approximation to the Poisson distribution suggests the measure $\mathrm{Chi}_{ij} = (N_{ij} - E_{ij})/\sqrt{E_{ij}}$, which is closely related to the Cochran (1954) and Mantel–Haenszel (1959) test statistics for testing independence in the c 2×2 tables formed by crossing $(A = i, A \neq i)$ with $(B = j, B \neq j)$ for each value of C. Test statistics for 2×2 independence ignoring C, such as

the Pearson chi-squared or the likelihood ratio chi-squared, would also be expected to behave similarly to (3). The latter, sometimes called the mutual information or entropy measure, was advocated for screening word collocation statistics by Dunning (1993) and compared to other measures in that context by DuMouchel et al. (1996). [But note that Church and Hanks (1991) used the term "mutual information" to refer to $\log_2(\mathrm{RR})$.] Using the formula for Poisson probabilities,

$$-\log_{10}(\Pr(X \geq N)) = -\log_{10}\left(\sum_{n \geq N} e^{-E}E^n/n!\right).$$

To avoid underflow in situations where $N \gg E$, the approximately equivalent quantity

$$\begin{aligned}-\log_{10}\left[\sum_{n=N}^{N+3} e^{-E}E^n/n!\right] &= E/\log(10) - N\log_{10}(E) \\ &+\log_{10}(N!) - \log_{10}\left[1 + E/(N+1)\right. \\ &\left.+E^2/(N+1)(N+2) + E^3/(N+1)(N+2)(N+3)\right]\end{aligned}$$

is used as a computational approximation when LogP becomes large (LogP > 12).

The concept behind such a measure is not that the null hypothesis is taken seriously, but only that the test statistic or its degree of significance might be a useful measure for ranking the degree of association among the different cells. For example, if $N_{ij} = 100$ and $E_{ij} = 1$, then $\mathrm{LogP}_{ij} = 158.4$, but of course a probability of $10^{-158.4}$ has no meaning except as a possible value for ranking cells in the table.

3.3 Empirical Bayes

The empirical Bayes approach tries to achieve the best of both of the previous approaches—to achieve the interpretability of the relative risk measures but also to adjust properly for sampling variation. Assume that each observed count N_{ij} is a draw from a Poisson distribution with unknown mean μ_{ij}, and that interest centers on the ratios $\lambda_{ij} = \mu_{ij}/E_{ij}$. But rather than treat the M values of λ_{ij} as unrelated constants, assume that each λ is drawn from a common prior distribution. This distribution is assumed be a mixture of two gamma distributions. The density function of a gamma distribution, having mean $= \alpha/\beta$ and variance $= \alpha/\beta^2$, is

$$g(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}/\Gamma(\alpha).$$

The prior probability density of λ is assumed to be

$$\begin{aligned}&\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) \\ &\qquad = Pg(\lambda; \alpha_1, \beta_1) + (1-P)g(\lambda; \alpha_2, \beta_2). \quad (4)\end{aligned}$$

Therefore the prior mean of λ under this mixture model is $P\alpha_1/\beta_1 + (1-P)\alpha_2/\beta_2$, and its prior variance is $P(1-P)(\alpha_1/\beta_1 - \alpha_2/\beta_2)^2 + P\alpha_1/\beta_1^2 + (1-P)\alpha_2/\beta_2^2$. The exact choice of prior distribution for λ is not so important as that it have several free parameters so that the distribution of λ can be fit using observed data. The density in (4) has five free parameters: $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, P)$. The family of gamma distributions are often used to model populations of Poisson rates because of the conjugate relationship between the Poisson and gamma distributions (Johnson and Kotz 1969; Bryck and Raudenbush 1992; O'Hagan 1994). The calculations are simplified in two respects: first, the marginal distribution of each N is a mixture of negative binomial distributions; and second, the posterior distribution of each λ is a mixture of two gamma distributions with modified parameters. Assuming that θ and E are known, then the distribution of N is

$$\begin{aligned}\Pr(N=n) &= Pf(n; \alpha_1, \beta_1, E) + (1-P)f(n; \alpha_2, \beta_2, E), \\ f(n; \alpha, \beta, E) &= (1+\beta/E)^{-n}(1+E/\beta)^{-\alpha} \\ &\quad \times \Gamma(\alpha+n)/\Gamma(\alpha)n!. \quad (5)\end{aligned}$$

Let Q_n be the posterior probability that λ came from the first component of the mixture, given $N = n$. From Bayes rule, the formula for Q_n is

$$\begin{aligned}Q_n = Pf(n; \alpha_1, \beta_1, E) \\ / \left[Pf(n; \alpha_1\beta_1, E) + (1-P)f(n; \alpha_2, \beta_2, E)\right]. \quad (6)\end{aligned}$$

The posterior distribution of λ, after observing $N = n$, can be represented as

$$\lambda | N = n \sim \pi(\lambda; \alpha_1 + n, \beta_1 + E, \alpha_2 + n, \beta_2 + E, Q_n), \quad (7)$$

where $\pi()$ is given by (4). Using well-known properties of gamma distributions, the posterior expectations of λ and of $\log(\lambda)$ are given by

$$\begin{aligned}E[\lambda | N = n] = Q_n(\alpha_1 + n)/(\beta_1 + E) \\ +(1-Q_n)(\alpha_2+n)/(\beta_2+E), \quad (8)\end{aligned}$$

and

$$\begin{aligned}E[\log(\lambda) | N = n] &= Q_n[\Psi(\alpha_1 + n) \\ &\quad - \log(\beta_1 + E)] + (1-Q_n)[\Psi(\alpha_2+n) \\ &\quad - \log(\beta_2 + E)], \quad (9)\end{aligned}$$

where $\Psi(x)$ is the digamma function, the derivative of $\log[\Gamma(x)]$. The empirical Bayes measure used to rank cell counts in this article is denoted EBlog2 and is defined as

$$\begin{aligned}\mathrm{EBlog2}_{ij} &= E[\log_2(\lambda_{ij}) | N_{ij}] \\ &= E[\log(\lambda) | N = N_{ij}]/\log(2), \quad (10)\end{aligned}$$

where the second expectation above is given by (9). The quantity EBlog2 is a Bayesian version of the information statistic $\log_2(\mathrm{RR})$, interpreted by Church and Hanks (1991) as the number of bits of information connecting row i and column j in the table of frequencies. Indeed, if E_{ij} and N_{ij}/E_{ij} are both large, then $\mathrm{EBlog2}_{ij}$ will approach $\log_2(\mathrm{RR}_{ij})$, as can be seen by representing $\log_2(\mathrm{RR}) = [\log(N) - \log(E)]/\log(2)$ and comparing this expression to Equations (8–10). For large arguments, $\Psi(x) \sim \log(x)$, and $\log(\alpha+N) - \log(\beta+E)$ will approach $\log(N/E) = \log(\mathrm{RR})$. However, when E or N/E are not large, then the effect of using EBlog2 is to "shrink" $\log_2(\mathrm{RR})$ toward smaller values, which is exactly the desired effect when sampling variation makes the true degree of association between $A = i$

and $B = j$ uncertain. To obtain a quantity on the same scale as RR, one can exponentiate EBlog2 to obtain

$$\text{EBGM}_{ij} = 2^{\text{EBlog2}_{ij}} \tag{11}$$

which is the geometric mean of the empirical Bayes posterior distribution of the "true" RR, and of course ranks the frequencies the same as Equation (10). Because of its easier interpretation, we present (11) in numerical examples.

To evaluate Equations (10) or (11), estimates of $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, P)$ are obtained by considering the marginal distribution of each N_{ij}, which is given in (5). The negative binomial distributions $f()$ in (5) are derived as a mixture of Poisson distributions, where the Poisson means have a gamma distribution (Johnson and Kotz 1969, p. 125). The likelihood function for θ is the product of M mixtures of two negative binomial densities:

$$L(\theta) = \Pi_{ij} \{Pf(N_{ij}; \alpha_1, \beta_1, E_{ij}) + (1-P)f(N_{ij}; \alpha_2, \beta_2, E_{ij}), \}. \tag{12}$$

The maximum likelihood estimate of θ is the vector that maximizes (12). The maximization involves an iterative search in the five-dimensional parameter space, where each iteration involves computing $\log[L(\theta)]$ and its first- and second-order derivatives. Since $\log[L(\theta)]$ is the sum of $M =$ several million terms, this computation must be approached delicately. In test runs using several datasets, the maximization typically takes between 5 and 15 iterations from the starting point $\theta = (\alpha_1 = .2, \beta_1 = .1, \alpha_2 = 2, \beta_2 = 4, P = 1/3)$. A weak rationale for these starting values goes as follows:

We assume that the majority of the cells (e.g., $1-P = 2/3$ of them) have values of λ clustered at or below the null hypothesis value of $\lambda = 1$. A corresponding gamma distribution with parameters $\alpha_2 = 2, \beta_2 = 4$ and having mean $= .5$ and standard deviation $= .35$ is suggested for this component. The remaining $P = 1/3$ of the cells, in which the chief dependencies reside, are hypothesized to have a λ-distribution with very high variance and a somewhat higher mean, with suggested parameters $\alpha_1 = .2, \beta_2 = .1$, having mean $= 2$ and standard deviation $= 4.5$. Note that our starting values have the property that the formula for the mean of λ, $P\alpha_1/\beta_1 + (1-P)\alpha_2\beta_2$, equals 1, which is reasonable since the Es are computed to have the same marginal totals as the Ns. However, except in choosing the starting point of the iterative search, the estimation algorithm enforces no such constraint.

4. EXAMPLE DATA

4.1 Data Selection From the FDA Spontaneous Reporting System Database

The objective of the analyses reported here is to show and compare methods and not to draw substantive conclusions. In particular, this analysis is a "big picture" analysis, with no restriction of included categories, even though it may make more sense scientifically and biostatistically to perform a more focused analysis. The work for this example starts with a tabulation of all combinations of stratum, drug, and event, together with N_{ijk}, a count of the number of reports in stratum k (based on 18 combinations of six time periods of about five years each and the three values of reported gender) that include drug i and event j. Originally, there were 7,695 drug codes and 1,234 event codes, with $\sum N_{ijk} = 4,922,802$. Note that the sum of all the counts is about 4.9 million, even though there are only about 1.2 million separate reports. This shows that most reports involve multiple drugs and/or events, and that therefore the counts represented by the N_{ijk} are clustered, a feature that was not taken into account by any of the models and methods of this article.

The data were slightly compressed to a total count of 4,864,480 as follows:

1. Collapse the 7,695 drug codes into 3,830 generic drug codes by trying to identify different drug codes that refer to the same chemical formulation. This is a difficult and time-consuming process with no guarantee that every such equivalence has been identified.
2. Retain only those 1,398 (generic) drugs involved in at least 100 reported combinations (i.e., $N_{i..} \geq 100$).
3. Retain only those 952 events involved in at least 100 reported combinations (i.e., $N_{.j.} \geq 100$).

Thus, $a = 1398, b = 952, c = 18$, and there are $M = ab = 1,330,896$ cells of interest. Using (1), E_{ij} was computed for the 385,734 drug-event combinations for which $N_{ij} > 0$. Thus, about 71% of the M cells are empty. In addition 137,051 other cells have $N_{ij} = 1$. At the other extreme, there are an 174 combinations having $N_{ij} > 1000$. (The largest value is $N = 7530$ for drug = LEVONORGESTREL [Norplant], event = METRORRHAGIA.) The 20 drugs having the most reported adverse event combinations and the five least reported drugs among those included, together with their total counts $N_{i..}$ are shown in Table 1.

The 20 most involved events and the five least involved events, with their total counts $N_{.j}$ are shown in Table 2.

Note that some of the drugs and events have slightly fewer than 100 reported combinations because the selection criterion for excluding rare drugs included counts for rare events and vice-versa.

4.2 Example Calculation of Baseline Frequencies and Rankings for a Single Drug and Three Adverse Events

Consider the most frequently reported drug in the SRS data base, FLUOXETINE [Prozac], and three different adverse events, chosen to include a very frequently reported event, HEADACHE, a moderately frequently reported event, AKATHISIA, and an infrequent event, POLYNEURITIS. Akathisia is a syndrome characterized by an inability to remain in a sitting posture, with motor restlessness and a feeling of muscular quivering, that sometimes appears as a side effect of antipsychotic and neuroleptic medication. Polyneuritis is a neurological syndrome marked by paresthesia of the limbs and muscular weakness or a flaccid

Table 1. The 20 Drugs With the Most Reported Adverse Event Combinations and the Five Least Reported Drugs Together With Their Total Couns $N_{i.}$

Drug (generic)	*FLUOXETINE*	*DIGOXIN*	*FUROSEMIDE*	*LEVONORGESTREL*	*RANITIDINE*	*DILTIAZEM*	
Example Brand	Prozac		Lasix	Norplant	Zantac	Cardizem	
Total Count	85304	62771	59364	52298	51621	48170	
Drug (generic)	*PREDNISONE*	*ALBUTEROL*	*ESTROGENS*	*ASPIRIN*	*IBUPROFEN*	*NAPROXEN*	*LEVOTHYROXINE*
Example Brand		Proventil	Premarin		Advil	Naprosyn	Levothroid
Total Count	45822	45414	45222	41875	41340	40654	40266
Drug (generic)	*CIMETIDINE*	*WARFARIN*	*NICOTINE*	*PHENYTOIN*	*ENALAPRIL*	*ALPRAZOLAM*	*ACETAMINOPHEN*
Example Brand	Tagamet	Coumadin		Dilantin	Vasotec	Xanax	Tylenol
Total Count	39992	39590	39203	38011	36703	36330	35654
Drug (generic)	*VITAMINS_W/IRON*	*FISH_OIL,_HYDROGENATED*	*CO-ADVIL*	*PHENIRAMINE*	*DELADUMONE_OB*		
Total Count	101	101	101	100	97		

paralysis, sometimes called Guillain–Barre syndrome. See PDR (1995). Table 3 shows statistics for these combinations taken from the subset of the SRS data described earlier.

The first four rows of Table 3 report the cell counts and marginal totals defining a 2×2 table for occurrence of a FLUOXETINE report and each of the three events. The next row shows the "naive" baseline frequency, computed by the familiar formula, row total times column total over grand total. For HEADACHE and AKATHISIA, this formula, which expects about 1.8% $(= 85,304/4,864,480)$ of every event type to involve FLUOXETINE, does not differ much from the adjusted baseline formula (1), perhaps because, although the SRS database goes back about 30 years, about three-quarters of all reported drug-event combinations came during the last 10 years, after PROZAC came on the market. But this is not true of POLYNEURITIS, which has been reported less often (perhaps only less often by that name) within the last 10 years. As a result, the stratified calculation of expected baseline frequency results in a number, 1.06, less than one-fourth of the unstratified result, 4.59, as shown in Table 3. Even though there were only three reports combining FLUOXETINE and POLYNEURITIS, the resulting relative report rate is $\mathrm{RR} = 3/1.06 = 2.84$, which is the 53rd largest value of RR among the 952 FLUOXETINE-event combinations. Although large in absolute value, allowance for normal Poisson variation deflates the significance of this result. The value of LogP from (3) is just 1.04, which ranks 163rd among the 952 FLUOXETINE-event values of LogP, and the empirical Bayes estimate EBGM in the last row of Table 3 suggests that 1.42 is a more accurate estimate of the "true" RR for this combination, ranking it 142nd by that measure.

The column of Table 3 for HEADACHE illustrates the opposite phenomenon of a fairly small value of $\mathrm{RR} = 1.23$, ranking 173 out of 952, but sample sizes are so large that it ranks 64th based on the Poisson significance level. However, in this case the empirical Bayes measure EBGM is virtually identical to the raw RR, and its ranking of 173 on that measure is equal to that of RR. Finally, the column for AKATHISIA in Table 3 provides an example in which both the sample sizes and the value of $\mathrm{RR} = 6.44$ are fairly large. In this case all three methods tend to agree, with RR and EBGM each ranking it 13th and LogP ranking it 19th. In summary, the example calculations in Table 3 show that relatively straightforward calculation of a relative rate or a Poisson significance test can provide some insight into drug-adverse event combinations in the SRS, even without reliable exposure data to use as a denominator, but that a more sophisticated methodology such as an empirical Bayes approach can help one negotiate the trade-off between being misled by sampling variation when sample sizes are small, and being misled by overreliance on significance probabilities when sample sizes are large. This latter type of error is especially likely when, as with the SRS, the data are subject to many types of reporting biases not at all accounted for by the significance probability calculation, so that values of RR as high as two or even higher might often be ascribed to such biases.

Table 2. The 20 Most Involved Events and the Five Least Involved Events, With Their Total Counts $N_{.j}$

Adverse Event	*RASH*	*NO_DRUG_EFFECT*	*PRURITUS*	*HEADACHE*	*DYSPNEA*	*URTICARIA*	*FEVER*	*DIZZINESS*
Total Count	145721	144756	85091	71209	68340	68191	67141	64842
Adverse Event	*NAUSEA*	*ASTHENIA*	*HYPOTENS*	*REACT_AGGRAV*	*PAIN_ABDO*	*PAIN*	*DIARRHEA*	*LEUKOPENIA*
Total Count	61520	55277	54554	54148	50731	47059	46196	44007
Adverse Event	*LIVER_FUNC_ABNORM*	*CONVULS*	*SOMNOLENCE*	*CONFUS*				
Total Count	43734	43434	42382	40523				
Adverse Event	*DEAF_PERM_TOTAL*	*ANEMIA_FOLIC_DEFIC*	*PIT_ACTIV_DEC*	*LYMPHANGITIS*	*LOW_BIRTH_WT*			
Total Count	101	100	99	98	97			

Table 3. Example Calculations for SRS Data Involving FLUOXETINE [Prozac] and Three Events

Event	HEADACHE	AKATHISIA	POLYNEURITIS
$N_{..}$ = total number of reported combinations	4,864,480	4,864,480	4,864,480
$N_{i.}$ = total number involving FLUOXETINE	85,304	85,304	85,304
$N_{.j.}$ = total number involving this event	71,209	3,001	262
N_{ij} = no. of reported FLUOXETINE-event combin.	1,614	328	3
$N_{i.}N_{.j.}/N_{..}$ = naive baseline frequency	1,249	52.6	4.59
E_{ij} = adjusted baseline frequency (Eq. (1))	1,309	51.0	1.06
$RR_{ij} = N_{ij}/E_{ij}$ [rank among 952 Events]	1.23 [173]	6.44 [13]	2.84 [53]
$LogP_{ij}$ (Eq. (3)) [rank]	15.9 [64]	146.5 [19]	1.04 [163]
$EBlog2_{ij}$ (eq. (10)*)	.301	2.68	.510
$EBGM_{ij} = 2^{EBlog2_{ij}}$ [Rank]	1.23 [173]	6.42 [13]	1.42 [142]

* Estimated θ is (α_1 = .2041, β_1 = .0582, α_2 = 1.4150, β_2 = 1.8380, P = .0969)

5. RESULTS

The presentation of results first focuses on which combinations of drug-event are selected by all three criteria, followed by a discussion of how these selections differ. Emphasis is on descriptive and statistical aspects of the results, rather than on possible medical or public health implications. In view of the limitations on interpretation of these data listed earlier, in this section the abbreviation "RR" will be understood to stand for "Relative Report Rate" rather than "Relative Risk." All results for the EBGM measure were computed using formulas (5)–(11), where the estimated hyperparameters (with estimated standard errors based on the inverse of the observed information matrix) are

	α_1	β_1	α_2	β_2	P
Estimate	.2041	.05816	1.4150	1.8380	.0969
St.Error	.0034	.00094	.0097	.0132	.0018

Note that the standard errors above are quite small, since the sample size is so large. The estimation results conform to the general rationale for the mixture model for λ presented at the end of Section 3: a majority component, having mean .77 and standard deviation .65, is contaminated by a distribution having mean 3.5 and standard deviation 7.8 in approximately 10% of the drug-event combinations. Figure 1 shows the estimated probability densities of λ and of $\log(\lambda)$; the two mixture components are the dashed and dotted curves in the figure.

5.1 Combinations Chosen by all Criteria

As a way of seeing what the different criteria have in common, Table 4 shows some statistics for the 65 combinations that ranked in the top 1,000 combinations (out of $M = 1.3$ million) by all three criteria. The fact that the intersection of the three sets of 1,000 had only 65 elements shows that there are real differences among the selection

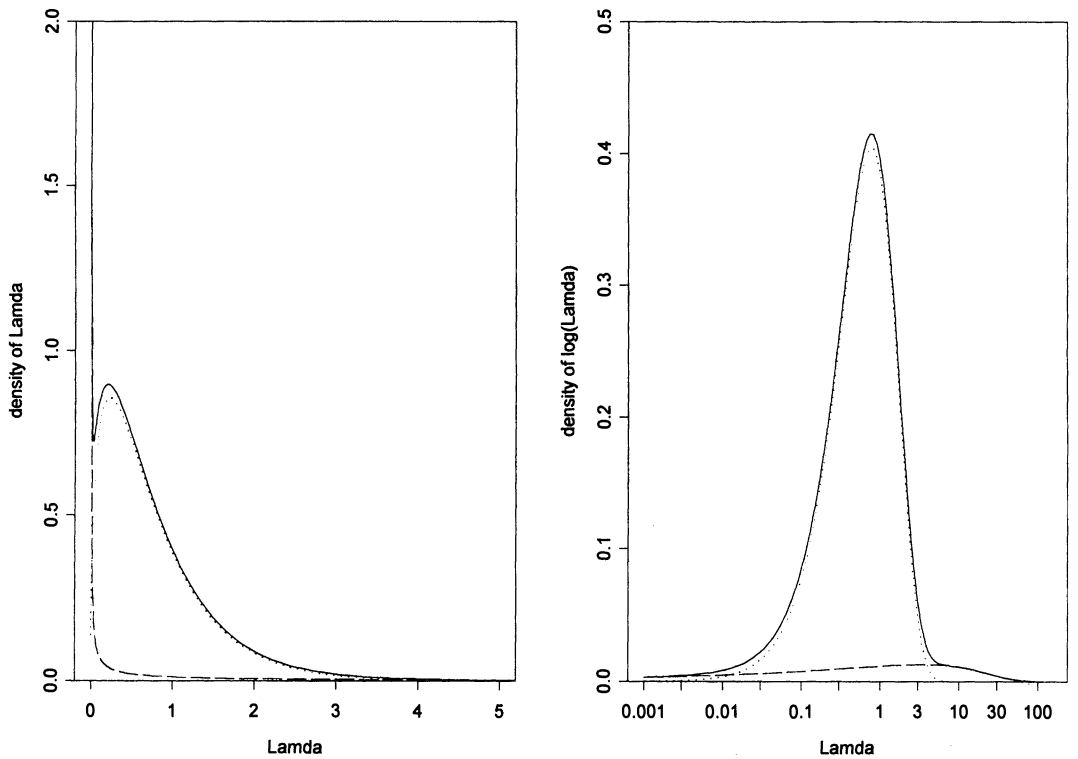

Figure 1. Estimated probability density functions of λ (left graph) and of $\log(\lambda)$ (right graph) for the SRS data. The solid curve is the total density, while the dashed curve is that of component 1 (9.7% probability) and the dotted curve is that of component 2 (90.3% probability).

criteria. (In fact, these 65 combinations are also the intersection of the two sets of top 1,000 chosen by the RR and LogP criteria; every combination in their intersection was also in the EBGM top 1000. This feature also held true when the subsets of top 10,000 and top 100,000 combinations were used instead.) Table 4 shows for each drug and event the values of N, E, RR, EBGM, and LogP. Next to each of the three criteria, in brackets, are the rankings of that combination by the corresponding criterion. The raw relative rate, RR $= N/E$, varies from 95 up to 2,695 in Table 4. The value of EBGM listed in the adjacent column is conceived as a "shrinkage" estimate of the "true" RR, and is always less than RR when RR is large. In general, the shrinkage formula for the empirical Bayes estimates, with the expression (9) involving $\log(\beta + E)$, result in a large amount of shrinkage whenever E is much less than $\min(\beta_1, \beta_2) = .0582$, as discussed earlier in Section 3.

Since the combinations in Table 4 were chosen to be among the most significant by all criteria, all of the Ns in this table are quite large and Poisson variation does not play much of a role here. Because of this, the EBGM measure is reasonably close to RR except when E is very small. This conservatism of EBGM could be viewed as being properly robust to unusually small baseline values that might themselves be unreliable. From the rankings in Table 4, it is quite striking how highly the empirical Bayes criterion ranks all the members of the intersection of the other two criterias' top 1,000 combinations. The EBGM top 11 are all included, as are 51 out of the top 100, and all 65 are in the top 181. The Bayesian criterion combines the requirements of practical significance and statistical significance in a single number.

It is encouraging to note that most of the combinations listed in Table 4 are medically unsurprising associations. Most are already described in standard references like PDR (1998), such as the dangers of infection from certain blood products, or of birth defects from certain drugs taken during pregnancy, or the discoloration of teeth due to certain antibiotics. Occasionally causation is reversed and a drug is associated with the disease for which it is often prescribed. For example GEMCITABINE HYDROCHLORIDE (Gemzar) is an antitumor agent used in cancer chemotherapy with no reported carcinogenic activity. Its presence on line 25 of Table 4 in combination with CARCINOMA GI is presumably such a reversal, providing an example of the need for expert interpretation and follow-up of these analyses.

The presence in Table 4 of adverse events associated with several vaccines provides an interesting example of the need to interpret the results cautiously. For example, consider line 43, the association of POLIOVIRUS VACCINE with SIDS, where $N = 146$, $E = 1.038$, RR $= 140$, EBGM $= 133$, LogP $= 252$. One certain source of bias is that the data are not adjusted for age of patient. The polio vaccine is given primarily to infants, and of course SIDS (sudden infant death syndrome) can only happen to infants; these facts alone would produce a substantial statistical association in an all-ages database. (A back-of-the-envelope calculation reveals that in the extreme case that all SIDS reports and all POLIO VACCINE reports came from infants, and if infants generate the proportion p of reported combinations, then the pooled-ages relative report rate for the POLIO VIRUS-SIDS combination would be RR $= 1/p$ even if there were no association within infant reports.) Although there is a field for age in the SRS database, difficulties with its format have so far delayed our ability to adjust for age as we did for gender and year of report. Another circumstance affecting the interpretation of this and other vaccine-related associations is the reporting requirement of the National Childhood Vaccine Injury Act, which requires that the manufacturer and lot number of vaccines administered to children be recorded by the health care provider in the vaccine recipient's permanent record, and the reporting of the occurrence, following immunization, of any event set forth in an official Vaccine Injury Table (which includes deaths). This law might result in a tendency for vaccine providers to have a larger report ratio of serious events such as SIDS to nonserious events such as rash or headache than would providers of other medications.

Table 5 helps to further explore the association of POLIO VACCINE and SIDS within our SRS methodology. Table 5a shows the results from all drugs having EBGM > 9 in association with SIDS, while Table 5b shows the results from all adverse events having EBGM > 9 in association with POLIO VACCINE. Considering Table 5a first, note that lines 1–4 and 9 all concern vaccines. (Lines 2–4 all refer to the same DTP vaccine; the fact that they show up separately in our analyses exposes a failure in our attempt to collapse results across all reports of the same drug by different names. On the other hand, the fact that the three values of EBGM are quite similar, and are more similar than are the three values of RR or the three values of LogP, provides some confirmation of the reliability of the SRS analysis in general and of the Bayesian methodology in particular.) This is consistent with our earlier speculation that vaccine-SIDS associations in general might be more likely to be reported. According to PDR (1998) there have been reports of SIDS following administration of DTP vaccine, although a large case-control study in the United States. (Hoffman et al. 1987) revealed no causal relationship between receipt of DTP vaccine and SIDS. The fact that the SIDS EBGM for polio vaccine is about twice as large as that for DTP vaccine must be balanced against the fact that the number of reported DTP-SIDS combinations is 251, versus only 146 POLIO VACCINE-SIDS reports, even though both vaccines have roughly similar administration rates. Of the other drugs in Table 5a, SULFISOXAZOLE and PEDIAZOLE are antibiotic formulations with similar ingredients and a total of 16 SIDS reports, while the other two drugs have just 4 and 3 SIDS reports, respectively. Turning now to Table 5b, which reports our analysis results about polio vaccine and various adverse events, the main thing to notice is that five of the 13 most associated events concern happenings at the injection site, even though polio vaccine is administered orally! Presumably these reports occur because the DTP vaccine or some other drug had been administered via injection at about the same time as the polio

Table 4. Combinations Appearing in the Top 1000 When Ranked by All Three Methods

	Drug	Event	N	E	RR[Rank]	EBGM[Rank]	LogP[Rank]
1	IOPHENDYLATE	ARACHNOIDITIS	293	.185	1583.73 [6]	1203.74 [1]	811.97 [74]
2	HEMOFIL	HIV TEST POS	306	.198	1543.39 [9]	1192.20 [2]	844.51 [69]
3	PERDIEM	STENO ESOPH	65	.024	2694.87 [3]	786.44 [3]	196.07 [491]
4	INPERSOL W/DEXTROSE	PERITONITIS	387	.542	714.05 [20]	644.36 [4]	938.25 [63]
5	PANALBA K.M.	DISCOLOR TOOTH	289	.425	680.75 [26]	598.12 [5]	695.04 [95]
6	DIANEAL	PERITONITIS	150	.228	658.94 [28]	523.82 [6]	359.27 [230]
7	ETIDOCAINE	TRISMUS	82	.105	780.57 [17]	500.62 [7]	202.97 [477]
8	BERACTANT	HEM LUNG	71	.089	797.58 [16]	480.41 [8]	176.55 [541]
9	OPCON A	MYDRIASIS	351	.717	489.53 [48]	452.43 [9]	793.66 [80]
10	DIETHYLSTILBESTROL	ANOMALY CONGEN UG	60	.093	647.58 [30]	395.90 [10]	143.95 [657]
11	CHOLINE	OPHTHALMITIS	58	.098	594.08 [35]	370.41 [11]	137.02 [689]
12	DURANEST W/EPINEPHRINE	TRISMUS	46	.073	628.21 [33]	347.88 [13]	110.00 [936]
13	DEMECLOCYCLINE	DISCOLOR TOOTH	257	.742	346.17 [91]	320.65 [16]	542.91 [127]
14	IMMUNE GLOBULIN, HUMAN	HEPATITIS C	218	.645	337.74 [94]	309.41 [17]	458.40 [164]
15	CORTISPORIN	PAIN EAR	132	.370	356.64 [81]	307.52 [19]	281.19 [319]
16	BSS PLUS	CORNEAL OPACITY	52	.129	402.19 [65]	275.84 [21]	114.16 [897]
17	DINOPROSTONE	UTER SPASM	56	.147	381.61 [72]	271.86 [22]	121.59 [818]
18	MIVACURIUM CHLORIDE	PARALYSIS FLACCID	104	.329	315.65 [108]	267.53 [24]	216.30 [446]
19	MYSTECLIN F	DISCOLOR TOOTH	134	.458	292.78 [130]	259.20 [25]	273.98 [337]
20	OXYTOCIN	HEM POSTPARTUM	103	.346	297.96 [124]	254.32 [26]	211.66 [456]
21	METHYSERGIDE	FIBRO RETROPERIT	87	.290	300.36 [122]	249.29 [27]	179.26 [536]
22	BSS	KERATITIS	74	.243	305.12 [118]	245.13 [28]	153.15 [620]
23	CHOLINE	CORNEAL OPACITY	77	.260	295.96 [129]	240.97 [29]	158.30 [603]
24	NONOXYNOL	BALANITIS	51	.166	306.82 [116]	225.98 [33]	106.01 [987]
25	GEMCITABINE HYDROCHLORIDE	CARCINOMA GI	225	.952	236.39 [200]	222.49 [34]	438.34 [177]
26	SODIUM HYALURONATE	KERATITIS	70	.274	255.20 [167]	209.67 [38]	139.52 [681]
27	DINOPROSTONE	LABOR ABNORM	231	1.052	219.50 [224]	207.74 [40]	442.58 [171]
28	TETRACYCLINE	ANOMALY TOOTH	380	1.776	213.96 [234]	207.02 [41]	722.95 [89]
29	BOTULINUM TOXIN A	PTOSIS	60	.238	252.56 [173]	201.89 [44]	119.47 [839]
30	DINOPROSTONE	FETAL DIS	167	.810	206.10 [252]	191.96 [49]	315.79 [271]
31	TICE BCG	GRANULOMA	55	.230	239.15 [192]	189.86 [50]	108.31 [956]
32	RIFABUTIN	UVEITIS	134	.665	201.45 [265]	184.84 [52]	252.31 [371]
33	MIVACURIUM CHLORIDE	INCREASED EFFECT	319	1.865	171.02 [356]	165.70 [60]	576.26 [120]
34	BENDECTIN	ANOMALY CONGEN	1106	6.785	163.00 [378]	161.57 [62]	1971.21 [19]
35	DORNASE ALFA	HEMOPTYSIS	64	.336	190.37 [301]	161.55 [63]	119.55 [838]
36	DIETHYLSTILBESTROL	ANOMALY CONGEN	2811	17.749	158.37 [398]	157.84 [65]	4972.33 [4]
37	FLUNISOLIDE	NASAL SEPTUM DIS	75	.431	174.04 [344]	152.74 [69]	137.00 [690]
38	CHOLINE	KERATITIS	116	.703	164.99 [372]	152.00 [70]	208.58 [464]
39	METFORMIN	ACIDOSIS LACTIC	309	2.074	149.02 [437]	144.82 [75]	539.88 [129]
40	BENDECTIN	ECTROMELIA	111	.721	154.02 [415]	142.14 [77]	196.35 [490]
41	GONADOTROPIN, CHORIONIC	OVAR DIS	86	.567	151.69 [423]	137.11 [80]	151.82 [627]
42	TETRACYCLINE	DISCOLOR TOOTH	1437	10.485	137.05 [510]	136.27 [81]	2453.16 [15]
43	POLIOVIRUS VACCINE, LIVE,	SIDS	146	1.038	140.60 [484]	132.87 [84]	252.13 [373]
44	SELENIUM SULFIDE	HAIR DISCOLOR	301	2.236	134.64 [526]	131.09 [87]	512.76 [141]
45	NONOXYNOL	CERVIX DIS	1074	8.161	131.61 [552]	130.64 [89]	1815.13 [22]
46	TICE BCG	CYSTITIS	69	.475	145.35 [457]	128.93 [94]	120.76 [827]
47	OPCON A	PAIN EYE	267	2.015	132.49 [547]	128.63 [95]	453.15 [166]
48	CEFOXITIN ENTEROCOL	PSEUDOMEM	138	1.016	135.79 [517]	128.16 [96]	236.31 [397]
49	TETRAHYDROZOLINE	PAIN EYE	187	1.430	130.77 [557]	125.46 [98]	316.73 [269]
50	MENOTROPINS	OVAR DIS	151	1.175	128.48 [583]	122.19 [99]	254.85 [366]
51	INFLUENZA VIRUS VACCINE	GUILLAIN BARRE SYND	62	.451	137.47 [507]	121.19 [100]	107.13 [975]
52	PILOCARPINE	MIOSIS	103	.812	126.83 [601]	118.02 [105]	173.66 [544]
53	PHENFORMIN	ACIDOSIS LACTIC	390	3.302	118.11 [681]	115.98 [109]	641.95 [103]
54	OXYTOCIN	FETAL DIS	95	.759	125.14 [617]	115.88 [110]	159.71 [598]
55	DIPHTHERIA-TETANUS TOXOID	SCREAMING SYND	147	1.220	120.47 [654]	114.76 [112]	244.06 [385]
56	OXYTETRACYCLINE	DISCOLOR TOOTH	74	.599	123.49 [633]	112.11 [118]	124.23 [793]
57	MITOMYCIN	UREMIA	218	1.995	109.29 [785]	106.05 [128]	352.17 [242]
58	POLIOVIRUS VACCINE, LIVE,	SCREAMING SYND	344	3.280	104.89 [831]	102.97 [135]	548.82 [125]
59	DTP VACCINE	SCREAMING SYND	966	9.643	100.18 [897]	99.55 [145]	1519.28 [29]
60	HAEMOPHILUS B POLYSACCHAR	MENINGITIS	176	1.711	102.86 [851]	99.31 [147]	279.98 [324]
61	COPPER	UTER DIS	4457	46.398	96.06 [975]	95.93 [154]	6922.91 [2]
62	OXYTOCIN	LABOR ABNORM	105	1.043	100.68 [887]	95.09 [161]	166.56 [571]
63	RUBELLA VIRUS VACCINE, LI	LYMPHADENO	115	1.162	98.97 [925]	94.01 [165]	181.47 [527]
64	ATRACURIUM BESYLATE	PARALYSIS FLACCID	100	1.051	95.13 [988]	89.88 [178]	156.25 [611]
65	PROPYLHEXEDRINE	DRUG DEPEND	72	.749	96.19 [968]	88.89 [181]	113.16 [909]

Table 5. Selected Results of Associations with POLIO VACCINE and SIDS

		N	E	RR	EBGM	LogP
	a) Drugs having EBGM > 9 in association with SIDS					
	Drug					
1	POLIOVIRUS VACCINE,	146	1.038	140.60	132.87	252.13
2	TRI-IMMUNOL	24	.258	93.03	74.99	38.02
3	DIPHTHERIA-TETANUS T	20	.272	73.44	59.63	29.80
4	DTP VACCINE	207	3.495	59.23	58.17	280.08
5	SULFISOXAZOLE	11	.134	81.86	55.62	17.24
6	RONDEC	3	.031	98.35	28.89	5.33
7	HYOSCYAMINE	4	.073	54.77	27.86	5.95
8	PEDIAZOLE	5	.123	40.70	25.88	6.68
9	BCG VACCINE	2	.019	104.31	13.50	3.74
	b) Adverse events having EBGM > 9 in association with POLIO VACCINE					
	Adverse Event					
1	SIDS	146	1.038	140.60	132.87	252.13
2	SCREAMING SYND	344	3.280	104.89	102.97	548.82
3	HYPOTONIA	116	2.130	54.47	52.89	153.36
4	PALLOR	85	3.094	27.47	26.87	88.08
5	INFLAM INJECT SITE	88	4.833	18.21	17.93	76.13
6	EDEMA INJECT SITE	77	4.971	15.49	15.25	61.67
7	MASS INJECT SITE	21	1.473	14.25	13.52	16.78
8	FEVER	751	63.973	11.74	11.72	506.73
9	MYELITIS	8	.607	13.19	11.33	6.58
10	ABSCESS INJECT SITE	15	1.244	12.06	11.27	11.20
11	OCULOGYRIC CRISIS	18	1.542	11.67	11.05	13.05
12	AGITATION	182	16.958	10.73	10.68	117.40
13	INJECT SITE REACT	138	14.810	9.32	9.26	81.69

vaccine. This shows how hard it can be to distinguish the effects of vaccines given in infancy based on a nonrandomized study. More generally, this discussion points out the need for expert interpretation and followup of any associations discovered while data mining, and of adjusting for variables like age that are known to have a strong effect on the responses of interest.

5.2 Combinations Uniquely Chosen by Each Criteria

To help us focus on how each criterion differs from the others, Table 6 lists combinations that are uniquely selected within the top 1,000 by each criterion. For example, the first 10 rows in Table 6 show combinations that are ranked in top 29 by the RR criterion, but that were all ranked worse than 1,000 by each of the other two criteria. Note that all of these combinations had $N \leq 2$ and $E < .0031$, defining a region of (N, E) that the RR criterion exclusively values. Moving down to the next block of rows in Table 6, the combinations exclusively chosen by EBGM, ranking between 205 and 246 by that criterion, but worse than 1,000 by the other methods, have quite a different distribution of (N, E). In this case N is moderate and E is small but not tiny, $.28 < E < .63$. Finally, the third block of rows in Table 6 shows 10 combinations ranked in the top 20 by the LogP criterion, but ranked worse than 1,000 by RR and EBGM. Here we see a preponderance of huge counts N, with large Es, $E > 193$, and relatively low values of N/E. The sizes of the uniquely chosen subsets from the sets of top 1,000 are 628, 504, and 811 for combinations chosen by RR, EBGM, and LogP, respectively. Table 6 presents the ten most highly ranked uniquely chosen combinations from among these sets. The combinations preferred by the empirical Bayes criterion represent a sensible tradeoff between large but extremely unreliable ratios, and extremely reliable but relatively small ratios.

To further illustrate the effect of the empirical Bayes estimation, Figure 2 shows, for the combinations used in Table 6, the observed values of RR with their associated 99.9% nonsimultaneous confidence intervals, based on the assumption of independent Poisson distributions for each N. Also plotted as open circles are the corresponding values of EBGM, showing how these estimates shrink towards smaller values of RR, with the degree of shrinkage depending on the accuracy of the corresponding statistic, as reflected by the widths of the confidence bars. Figure 2 shows clearly that the middle group of drug-event combinations is estimated to have the largest set of λs, and that the first and third groups are estimated to be about the same on average, even though the first group has values of RR that are hundreds of times greater.

6. DISCUSSION

This article argues that a new empirical Bayes screening criterion, a further development of the one presented in DuMouchel et al. (1996), can be generally useful for the data mining task of screening very large, sparse frequency tables for cells showing association. Measures based on raw ratios of observed to expected seem too ready to select cells with very small counts. One way to reduce the sampling variation in RR would be to require a minimum value of N, such as at least 5 observations in a cell, before including it in the ranking according to RR. But one needs a rationale for choosing the cutoff value. And even after choosing 5,

Table 6. The Ten Most Highly Ranked Uniquely Chosen Combinations for Each Method, Based on Choosing Each Method's Top 1,000 Combinations

Method:Rank	Drug	Event	N	E	RR	EBGM	LogP
RR:1	NISOLDIPINE	HEPATITIS NONSPECIFIC	1	.00028	3561.9	2.4	3.6
RR:2	ACARBOSE	HEPATITIS NONSPECIFIC	1	.00036	2817.0	2.4	3.4
RR:5	ORTHO-NOVUM 1/80	CARCINOMA LIVER	1	.00054	1840.0	2.4	3.3
RR:10	PLATELET CONCENTRATE, HUMAN	LIVER DAMAGE AGGRAV	1	.00076	1320.9	2.4	3.1
RR:11	URSODIOL	LIVER DAMAGE AGGRAV	1	.00084	1187.5	2.4	3.1
RR:12	EMLA	HYPALGESIA	1	.00090	1114.1	2.4	3.0
RR:19	ORTHO-NOVUM SQ	CARCINOMA LARYNX	1	.00139	720.7	2.4	2.9
RR:22	COLFOSCERIL PALMITATE	INTEST SMALL PER	2	.00286	699.1	19.7	5.4
RR:23	AVC	BALANITIS	1	.00144	694.6	2.4	2.8
RR:29	HYDROCORTISONE-NEOMYCIN-POLYM	OTITIS EXT	2	.00308	649.0	19.6	5.3
EBGM:205	GANCICLOVIR	RETINITIS	57	.62521	91.2	83.0	88.5
EBGM:214	GLYCINE	HYPONATREM	44	.47895	91.9	81.4	68.7
EBGM:216	SELENIUM SULFIDE	SEBORRHEA	43	.46996	91.5	80.9	67.1
EBGM:220	SUPROFEN	PAIN KIDNEY	37	.40002	92.5	80.1	58.0
EBGM:226	NONOXYNOL	PENIS DIS	44	.49613	88.7	78.8	68.0
EBGM:232	MINIZIDE	HYPOKALEM	33	.35936	91.8	78.3	51.8
EBGM:234	PRILOCAINE	METHEMOGLOBIN	44	.50304	87.5	77.9	67.8
EBGM:236	FENFLURAMINE	HYPERTENS PULM	27	.28538	94.6	77.7	42.9
EBGM:240	PENTAZOCINE	FIBRO INJECT SITE	36	.40419	89.1	77.2	55.9
EBGM:246	TROPICAMIDE	MYDRIASIS	27	.29057	92.9	76.6	42.6
LogP:3	LEVONORGESTREL	METRORRHAGIA	7530	524.28003	14.4	14.4	5673.7
LogP:5	LEVONORGESTREL	REACT UNEVAL	5043	259.50101	19.4	19.4	4422.9
LogP:6	WARFARIN	PROTHROMBIN DEC	4355	193.60800	22.5	22.5	4083.2
LogP:8	NICOTINE	APPLICAT SITE REACT	5085	358.59900	14.2	14.2	3805.9
LogP:9	DIATRIZOIC ACID	URTICARIA	5404	442.64801	12.2	12.2	3719.8
LogP:10	MINOXIDIL	ALOPECIA	4479	262.60800	17.1	17.0	3688.6
LogP:11	PERMETHRIN	NO DRUG EFFECT	4029	249.04201	16.2	16.2	3231.3
LogP:14	INSULIN HUMAN	NO DRUG EFFECT	6194	1003.65002	6.2	6.2	2643.8
LogP:16	ESTRADIOL	APPLICAT SITE REACT	3664	325.30600	11.3	11.3	2405.5
LogP:20	IOTHALAMIC ACID	URTICARIA	2924	248.90401	11.8	11.7	1968.8

say, as the threshold, the high values of RR are likely to cluster at $N = 5$, and the even higher values of RR left out having $N = 4$ will always be tempting. At the opposite extreme, measures based on significance testing focuses too heavily on cells with very large counts, even if the actual ratio of N/E is not very impressive. Besides avoiding these problems, the EBGM measure has two positive advantages. First, the estimation of the hyperparameters allows the ranking of cells to adapt to the specific population of interest. With the other measures, if (N_1, E_1) ranks higher than (N_2, E_2) in one database, it ranks higher in all databases. With the empirical Bayes measures, the ranking may reverse depending on the distribution of the M observed (N, E) pairs. Second, because of the term $\log(\beta+E)$ in (9), EBGM is robust to the accidental presence of very tiny values of E that can inflate the other criteria.

Although the empirical Bayes estimation procedure can always fit the distribution of λ at a gross level, so that the estimated mean and variance of λ are approximately right, there is evidence in these data that the true distribution of λ is not exactly as assumed. For example, the fitted mixture model parameters lead to a calculation that $P(\lambda > 200) = 2.6 \times 10^{-8}$, so that the expected number of times that $\lambda > 200$ out of $M = 1{,}330{,}896$ drug-event combinations is about .03. But actually 44 combinations have the estimator EBGM > 200. The true distribution of λ seems to have even more outliers than the mixture of gamma distributions allows, at least if we assume that the E_{ij} are estimated without error. In view of this lack of fit to the model, it is best not to overinterpret the meaning of our two-component prior distribution. Rather than identifying two distinct populations of drug-event combinations, the mixture distribution is more likely just a somewhat arbitrary five-parameter distribution useful in locating the largest λs for screening purposes. But the gamma mixture model fits very much better than a single gamma distribution—the estimate of (α_1, β_1) is over 100 standard errors away from that of (α_2, β_2) in both dimensions.

More complex procedures for dealing with the prior distribution $\pi(\lambda)$ include using more than two gamma mixture components, using a nonparametric estimate for $\pi(\lambda)$, or using methods specifically oriented toward ranking the unknown λs, such as the methods described by Laird and Louis (1989), or Shen and Louis (1998). But note that our setup differs from that of those two articles in the extra complication of differing baseline frequencies, making the marginal distributions of the observed frequencies nonidentical. As a less complex procedure, one might just assume that the prior distribution is the exponential distribution, $\pi(\lambda) = e^{-\lambda}$, which is the simplest gamma distribution having expectation 1, and equivalent to taking $P = \alpha_1 = \beta_1 = 1$ in (4). This leads to the simple posterior expectation $E(\lambda_{ij}|N_{ij}) = (N_{ij}+1)/(E_{ij}+1)$, which, as a data mining criterion, has lower variance than RR and more validity than LogP, especially for larger values of E_{ij}. Making such an extreme unchecked assumption would

prompt howls of outrage from many readers of this journal, but the resulting estimates might work well for the purpose of screening drug-event combinations. After the present article was submitted, the author was made aware of the publication of "A Bayesian Neural Network Method for Adverse Drug Reaction Signal Generation," (Bate et al. 1998) which presents a method very similar in spirit to the above suggestion of using a simple exponential prior distribution fixed in advance without empirical Bayes estimation of hyperparameters. Bate et al. (1998) use separate beta-binomial Bayesian estimates for $P(\text{drug} = i)$, $P(\text{event} = j)$, and $P(\text{drug} = i, \text{event} = j)$. Their choices of beta hyperparameters are chosen to shrink the estimated probabilities toward the baseline hypothesis of independence of drug and event, and the strength of the shrinkage is approximately that which would result from the use of $\pi(\lambda) = e^{-\lambda}$ in the gamma-Poisson model. Using normal and delta approximations to their beta posterior distributions, they estimate the posterior mean and variance of $\text{IC} = \log_2[P(\text{drug} = i, \text{event} = j)/P(\text{drug} = i)P(\text{event} = j)]$, which are somewhat complicated formulas but can be shown to be approximately equal to $\log_2[(N_{ij}+1)/(E_{ij}+1)]$ and $\log_2(e)^2/(N_{ij}+1)$, at least when E_{ij} is not too small and the unstratified $(c = 1)$ definition of baseline frequency is used. Bate et al. (1998) analyze the World Health Organization database of adverse drug reactions held by the Uppsala Monitoring Centre in Sweden, rather than the U.S. FDA SRS database. They do not compare the results from their method, which they call a Bayesian confidence propagation neural network (BCPNN) to other measures, nor do they examine the goodness of fit of their prior distribution assumptions. The description in Bate et al. (1998) does not explain why their Bayesian calculations are described as a neural network. Their analyses of the WHO database claim to find several previously unreported adverse drug reactions.

The results of the present article echo those in DuMouchel et al. (1996), in which an empirical Bayes measure compared favorably, applied to the natural language processing domain, to two other commonly used measures, the mutual information measure advocated by Dunning (1993) and a conditional probability measure used by Freidman et al. (1995). The methodology presented here advances that in DuMouchel et al. (1996) in two respects. First, a more general multidimensional framework is presented here, whereas the earlier work was restricted to analysis of a two-way table. The notation and example of this paper focuses on the situation where a third dimension is used only for stratification but is not of primary interest; however the extension to more general situations should be clear. Second, this article achieves a better fit to the data by introducing a five-parameter mixture of gammas as the prior distribution for λ, whereas the previous work used a two-parameter gamma distribution. One disadvantage of the use of a gamma mixture instead of a single gamma is that posterior percentiles for λ no longer have a simple formula in terms of the percentiles of a chi-squared distribution, as was provided in DuMouchel et al. (1996). Although a computer program to calculate posterior distribution percentiles for our mixture model would not be a big project, we have so far relied on the quick-and-dirty approximate 95% posterior probability interval

$$\text{EBGM}_{ij} \exp\{-2/\sqrt{(N_{ij}+1)}\} < \lambda_{ij} < \text{EBGM}_{ij} \exp\{2/\sqrt{(N_{ij}+1)}\}. \quad (13)$$

Equation (13) is based on a log-normal approximation and uses $1/\sqrt{(N_{ij}+1)}$ as the posterior coefficient of variation of λ, which is what results from use of the simple exponential prior distribution for λ discussed earlier.

In the previous work, a cursory examination of the uniquely chosen word pairs clearly showed that those chosen by the empirical Bayes method were more "interesting" from a substantial point of view. For the analysis of the SRS database, work is ongoing with medical collaborators in which preliminary results are showing new and potentially important relationships. Some features of the SRS data base as currently available hinder making best use of it. For example, as discussed above, stratifying on age of patient would certainly make sense, but the age variable in the SRS data is not available in a consistent format and needs further preprocessing before it can be useful. Another problem is the fact that the names of drugs and adverse events are not fully standardized, and our attempt to collapse over drugs having multiple names was not completely successful.

An implementation question is how many categories of A and B to include in the overall analysis. One could reduce a and/or b either by restricting attention to a subset of categories or by pooling categories. Assuming that the larger value of $M = ab$ does not present a computational problem, are there statistical or substantial advantages to lowering the number of cells? Because of the statistical requirement discussed earlier that the marginal totals $N_{i.}$ and $N_{.j}$ should all be large, it can often help to pool substantially related rare categories, assuming no violence is being done thereby to interpretational objectives. Dropping categories and analyzing only a subset of the database is obviously appropriate if it is desired to narrow the objective of the data mining project. For example, in the analysis of adverse event-drug reports, one might only be interested in psycho-active drugs and/or in adverse events involving a particular system like the cardiovascular or nervous system. Within such a reduced and more homogeneous database, the baseline frequencies will usually better fit the raw counts, so that comparisons of N_{ij} to E_{ij} require different interpretations—perhaps a lower threshold for being "interesting."

The empirical Bayes model used here specifically assumes that the ratio $\lambda = \mu/E$, where $\mu = E[N]$, is of primary interest. Although this is often true, in some domains the difference $\mu - E$ may be more important to estimate accurately—for example, if the difference corresponds to a number of lives lost or a potential dollar profit. In such situations an empirical Bayes model for the difference can be constructed, as was done by DuMouchel (1983), which

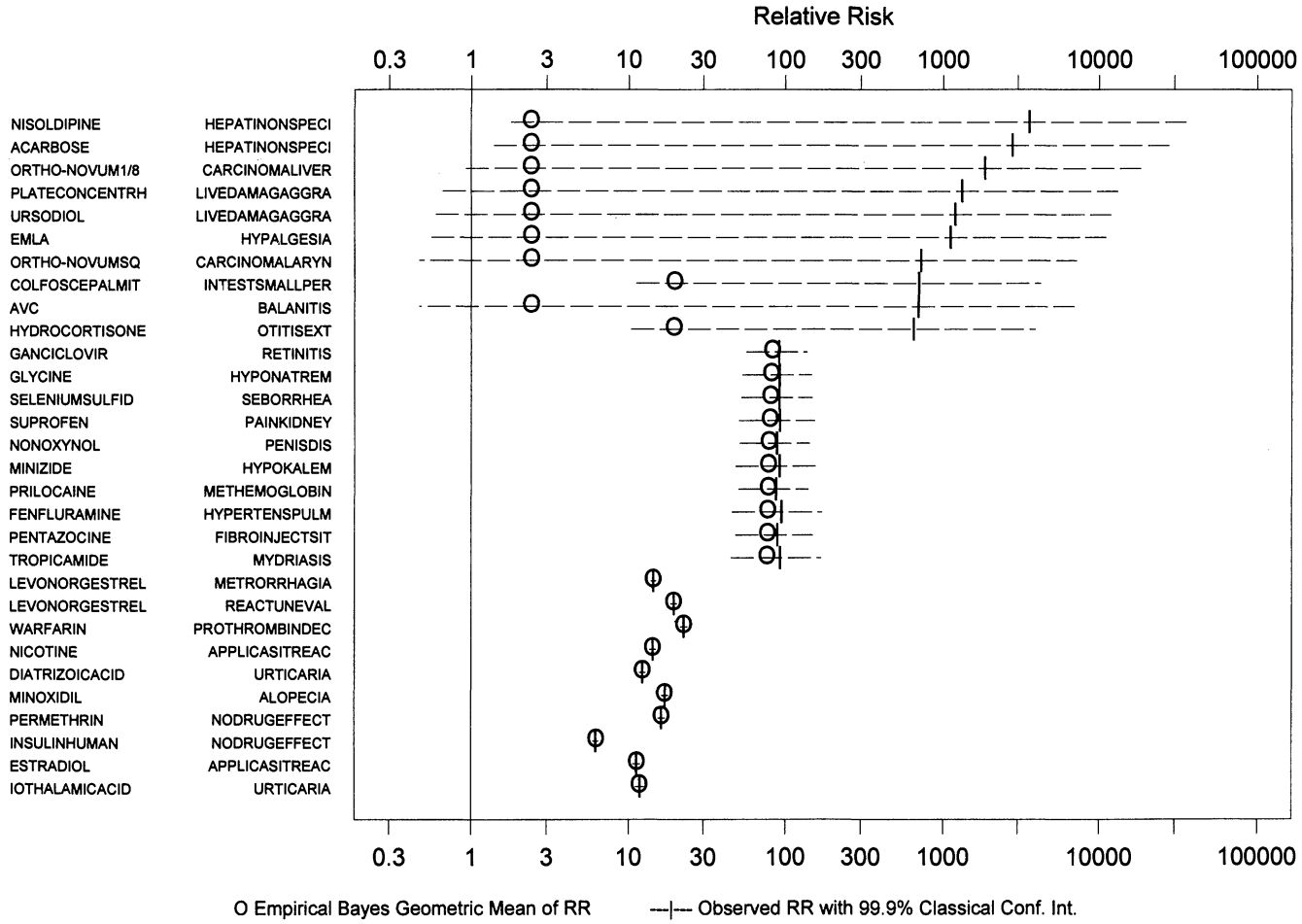

Figure 2. Observed Relative Report Rates with 99.9% Confidence Limits and Bayesian Estimates. (Data from Table 4, top ten combinations chosen exclusively by each criterion).

estimated claims frequencies in the context of setting automobile insurance rates.

The use of Bayes or empirical Bayes methods to estimate cell probabilities in a frequency table has a long history. See, for example, Bishop, Fienberg, and Holland (1975, chap. 12) for an extensive discussion and many references. The original contribution of the present work is the focus on a data mining goal with the ability to scale up to examples with millions of cells, and the comparison to classical methods for ranking deviations of cells from a baseline model. Much of the previous work uses a Dirichlet-multinomial model rather than the gamma-Poisson model. The gamma-Poisson formulation is practically equivalent and involves simpler calculations. This research also has much in common with work oriented towards defining residuals for models involving categorical data. See, for example, Edwards (1995b) or Agresti (1990). But model criticism is not really an issue here, since it is taken for granted that the baseline model is false. Rather, the focus is on pragmatic ways to rank cell deviations because the deviations themselves are of interest, at least the ones where $N \gg E$. And unlike the typical scenario where residuals are tested for significance, estimation of λ is the real goal. Work similar in spirit to the present article is also described in the literature on empirical Bayes adjustment of disease rates or other small-area estimation problems, as in Devine, Halloran, and Louis (1994).

Future work will present examples and extensions geared to search for "interesting" cells in higher dimensional tables. Two such extensions in the context of analyses of the SRS database are the identification of gender differences in adverse drug reactions and the identification of adverse reactions due to the interaction of two or more drugs. A program named GPS (Gamma-Poisson Shrinker) for performing the likelihood function maximization and computing the empirical Bayes criteria is available. A compressed Winzip file containing the program and help files, which run on a PC under the Microsoft Windows operating system, may be obtained from the Internet location ftp://ftp.research.att.com/dist/gps/.

[Received July 1998. Revised April 1999.]

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

Almond, R. G. (1995), *Graphical Belief Modeling*, New York: Chapman and Hall.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., and DeFreitas, R. M. (1998), "A Bayesian Neural Network Method for Adverse Drug Reaction Signal Generation," *European Journal Clinical Pharmacology*, 54, 315–321.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Bryck, A., and Raudenbush, S. (1992), *Hierarchical Linear Models*, Newbury Park, CA: Sage Publications.

Church, K., and Hanks, P. (1991), "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 16, 22–29.

Cochran, W. G. (1954), "Some Methods of Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417–451.

Devine, O. J., Halloran, M. E., and Louis, T. A. (1994), "Empirical Bayes Methods for Stabilizing Incidence Rates Prior to Mapping," *Epidemiology*, 5, 622–630.

DuMouchel, W. (1983), "The 1982 Massachusetts Automobile Insurance Classification Scheme," *The Statistician*, 32, 69–81.

DuMouchel, W., Friedman, C., Hripcsak, G., Johnson, S., and Clayton, P. (1996), "Two Applications of Statistical Modeling to Natural Language Processing," *AI and Statistics V*, eds. D. Fisher and H. Lenz, Springer-Verlag.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, 19, 61–74.

Edwards, D. (1995a), *Introduction to Graphical Modeling*, New York: Springer-Verlag.

——— (1995b), "Residual Analysis in Undirected Graphical Models," in *Bulletin of the International Statistical Institute, Proceedings of the 50th Session*, Book 1, pp 431–440.

Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S., and Clayton, P. (1995), "Natural Language Processing in an Operational Clinical Information System," *Natural Language Engineering*, 1, 1–28.

Hand, D. J. (1998), "Data Mining: Statistics and More?" *The American Statistician*, 52, 112–118.

Hoffman, H. J., Hunter, J. C., Damus, K., Pakter, J., Peterson, D. R., van Belle, G., and Hasselmeyer, E. G. (1987), "Diptheria-Tetanus-Pertussis Immunization and Sudden Infant Death; Results of the National Institute of Child Health and Human Development Cooperative Study of Sudden Infant Death Syndrome Risk Factors," *Pediatrics* , 79, 598–611.

Johnson, N., and Kotz, S. (1969), *Discrete Distributions*, Houghton Mifflin, now distributed by New York: John Wiley.

Laird, N. M., and Louis, T. A. (1989), "Empirical Bayes Ranking Methods," *Journal of Educational Statistics*, 14, 29–46.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Oxford University Press.

Mantel, N., and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719–748.

Miller, R. G. (1966), *Simultaneous Statistical Inference*, New York: McGraw Hill.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics, vol. 2B, Bayesian Inference*, New York: Halstead Press (Wiley).

PDR (1995), *PDR Medical Dictionary* (1st ed.), Montvale, NJ: Medical Economics.

——— (1998), *Physicians' Desk Reference* (52nd ed.), Montvale, NJ: Medical Economics.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann.

Shen, W., and Louis, T.A. (1998), "Triple-Goal Estimates in Two-Stage Hierarchical Models," *Journal of the Royal Statistical Society*, Ser. B, 60, 455–471.

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B., 13, 238–241.

Stati cal Sciences (1995), *S-Plus Guide to Statistical and Mathematical An ysis*, (ver. 3.3), Seattle, WA: StatSci, a Division of MathSoft.

Discussion

Robert T. O'Neill and Ana Szarfman

We congratulate William DuMouchel for developing an innovative and exciting methodological approach for screening a large database of adverse event reports that spans 30 years of cumulative information. DuMouchel's analysis of a large frequency table by means of a statistical model employing Bayesian shrinkage estimators provides an efficient way to systematically screen a large structured database and enhance its use for identification and surveillance of both known and unknown drug-adverse event associations. In this commentary, we acknowledge that DuMouchel's work was carried out in response to a need identified by the Food and Drug Administration (FDA) and supported under a grant from FDA's Office of Women's Health. Our commentary on DuMouchel's article will cover a range of issues including the structure and content of the FDA database, the relevance of DuMouchel's model formulation to the problem of alerting for drug-event associations; some desirable features the model should account for; the role of visual graphics in helping to interpret the empirical Bayes scores; some related work on signal generation; and finally some ideas for future developments and refinements to enhance the utility of the methods.

Hand (1998) described data mining as "a new discipline, lying at the interface of statistics, database technology, pattern recognition, and machine learning, and concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners." In this sense, the problem addressed by DuMouchel is motivated by the need to develop an exploratory approach to screen a large database of adverse event reports that is fast and efficient, that is able to evaluate all drugs and adverse events in the database, and that is capable of examining any portion of the database using successively more detailed follow-up queries to explore the strength of signals. The goal is to both describe the database in a big picture manner and to be able to proceed from this big picture down to the particulars in a logical manner that would facilitate medical understanding and insight into the drug-event associations in the database. DuMouchel formulated an approach to meet this goal by statistical modeling of a large database with sparse cells and we have further combined these methods with visual graphics and pattern recognition strategies to serve as a tool for

Robert T. O'Neill is Director, and Ana Szarfman is Medical Officer, Office of Biostatistics, Food and Drug Administration, Center for Drug Evaluation and Research, 5600 Fishers Lane, 15B-45, Rockville, MD 20857 (Email: oneill@cder.fda.gov). Ana Szarfman is the principal investigator at FDA for this grant. Disclaimer: The opinion expressed in this article are the professional views of the authors and do not necessarily reflect the official position of the United States Food and Drug Administration. The authors thank the following members of the CrossGraphs team for their support of our data mining and data visualization activities: David Fram, currently affiliated with Lincoln Technologies, Inc.; Jeremy Pool, Ava-Robin Cohen, and Ilya Yunus from Belmont Research; and Jay Levine of FDA.