MULTIVARIATE MIXTURE MODELS FOR CLASSIFICATION OF ANEMIAS

Christine E. McLaren[1], Igor V. Cadez[1], Padhraic Smyth[1],
and Geoffrey J. McLachlan[2]
[1]University of California, Irvine; [2]The University of Queensland
C. E. McLaren, Epidemiology Division, 224 Irvine Hall,
University of California, Irvine, CA 92697

KEY WORDS: EM algorithm, Probabilistic Clustering, Multivariate mixtures, Anemia

ABSTRACT

Over one billion people in the world are anemic and at risk for major liabilities. Previous models proposed to differentiate between disorders of anemia on the basis of red blood cell measurements have been limited by the need to use printed output from automated blood sample analyses. We developed electronic methods to capture multivariate red cell data measured by flow cytometric blood cell counting instruments and devised a general bilevel framework for classification that includes (1) fitting mixture densities to the multivariate grouped and truncated distribution for each individual, and (2) discrimination between patient subgroups on the basis of distribution parameter estimates. Classification by fitting normal density models, with leave-one-out cross validation, achieved 97% and 99% correct classification for controls and patients, respectively.

1. INTRODUCTION

According to population estimates, over one billion people in the world have anemia, defined as a reduction in the circulating red cell mass that may diminish the oxygen-carrying capacity of the blood. Iron deficiency anemia, attributed to an imbalance between dietary iron supply and physiological requirements for growth and reproduction, is the most common nutritional anemia [1]. Major liabilities including mental and motor developmental defects in infants [2], and weakness, weight loss, and impaired work performance in adults [3]. Other nutritional anemias include vitamin B_{12} deficiency and folate deficiency. The anemia of chronic disorders found in infectious diseases such as tuberculosis, typoid, and smallpox and in noninfectious disorders including rheumatoid arthritis, Hodgkin disease, metastatic carcinoma, is usually moderate and rarely symptomatic, while thalassemia, a form of severe anemia caused by mutations (or deletions) in or around the globin chain DNA and accompanied by a disturbance of hemoglobin synthesis, may lead to organ damage and premature death. Chronic alcohol ingestion is often associated by anemia as a result of poor nutrition, gastrointestinal bleeding, or the toxic efffect of alcolohol on the production of erythrocytes. Alcoholics may also develop coincident iron deficiency and folate deficiency [4].

Flow cytometric blood cell counting instruments make measurements on *each* red cell using a laser light scattering system. This technology provides the red cell volume distribution, hemoglobin concentration distribution, and the joint red cell volume and hemoglobin concentration distribution. Since different causes of anemia may result in characteristic alterations in these distributions, we hypothesized that classification based on modeling of the multivariate distribution of red cell volume and hemoglobin concentration would be useful for diagnostic evaluation of anemia. Our study is the first to model and classify these multivariate distributions.

Methods have been developed for detection of two-component mixtures of lognormal distributions and utilized to characterize and quantify subpopulations of red blood cells in developing iron deficiency anemia and subsequent treatment for the disease [5 – 7]. While these methods have been applied to analysis of univariate red blood cell volume distributions, no suitable statistical methods are currently available for analysis of multivariate distributions arising from multiple measurements made on a single blood cell, such as the volume and hemoglobin concentration of a red blood cell.

We now describe a general framework that includes the following: (1) development of techniques to model multivariate mixtures of distributions from grouped and truncated data, (2) description of the

Reprinted from the 2000 Proceedings of the Biometrics Section

bivariate distribution of red cell volume and hemoglobin concentration in patients with anemia and controls, and (3) classification of patient subgroups on the basis of distribution parameter estimates. Analysis of data from 90 healthy individuals and 146 patients with documented disorders of anemia showed that mixture modeling on parameter estimates with leave-one-out cross validation achieved 97% and 99% correct classification for controls and patients, respectively. We conclude that these methods provide a means for automated screening for disorders of anemia and monitoring the response to therapy.

2. METHODS

2.1 Patients and Reference Group

This study was performed at the Western Infirmary, Glasgow, Scotland after Institutional Review Board approval was obtained. We collected blood samples from a reference group of healthy individuals and patients with documented disorders of anemia. Diagnoses and body iron status were confirmed by examination of blood films, iron studies, and red cell indices. Reference ranges were as follows: hemoglobin (HGB) 13.5-17.5 g/dL (males), 12-16 g/dL (females); mean cell hemoglobin concentration (MCHC) 33.4-35.3 g/dL; and mean cell volume (MCV) 80-100 fl. We analyzed data from 90 healthy individuals and 146 patients. Patients were divided into two subgroups, those with microcytosis including iron deficiency anemia (n=82), thalassemia (n=8) and anemia of chronic disease (n=16), and those with macrocytosis including vitamin either B_{12}/folate deficiency (n=12), and alcoholic liver disease (n=28). For blood cell analysis, we used a flow cytometric blood cell counting instrument Technicon H*1 (Bayer Diagnostics, Tarrytown, New York, USA). For each sample, measured in duplicate, the data consisted of a cytogram, i.e. bivariate histogram, with a range of 0 to 200 fl for cell volume and 0 to 50 g/dl for hemoglobin concentration.

2.2 Mixture Modeling

We developed techniques to model the joint distribution of red cell volume and hemoglobin concentration as a mixture of two multivariate lognormal distributions. Finite mixture models have been fit to univariate grouped and truncated data by maximum likelihood via the Expectation-Maximization (EM) algorithm [5, 8–9]. For our studies, the EM algorithm was extended to evaluate multidimensional integrals over two-dimensional regions and numerical integration techniques were employed to improve computational efficiency [See Cadez et al. [10] and the Appendix). For analysis of data from healthy individuals and patients, we developed a new bilevel modeling technique as described in the Appendix. Results from initial analyses of data from controls and patients with iron deficiency anemia are reported elsewhere [11]. In brief, we first identified mixtures of two subpopulations of cells within a single blood sample by fitting a two-component lognormal mixture model to each individual distribution. The parameter estimates from each fitted distribution were recorded. These included the mixing weight for the larger proportion, volume and hemoglobin concentration means and variances for each component, the estimated correlation between volume and hemoglobin concentration for each component, and covariances. Second, for discrimination between patient subgroups, a supervised classifier for the parameter sets of all subjects from the same disease subgroup (control, microcytic anemia, macrocytic anemia) was used to fit a normal density model to each group.

2.3 Classification

Classification was performed using Bayes rule for the posterior class distribution:

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{C} \qquad (1)$$

where $p(x|c_i)$ is a probability density function under the i-th single normal subpopulation, $p(c_i)$ is the prior for the disease (i.e. the ratio of patients in the i-th group to the total number of patients), and C is a normalizing constant. Both $p(x|c_i)$ and $p(c_i)$ are estimated from the data as described in the Appendix. To estimate the overall accuracy of discrimination of patient subgroups, we used leave-one-out cross validation. For this type of cross validation we removed a data point from the data set and trained the three single normal density models. The excluded data point was then evaluated using Bayes rule and assigned to a class. This process was repeated for each of the data points in the data set. The overall percent of correctly classified distributions was calculated.

3. RESULTS

3.1 Distribution Modeling and Classification

Figure 1 shows histograms from two representative subjects. Each distribution represented about 40,000 red blood cells measured on a single blood sample. Figure 1A shows the distribution from a healthy male with estimated geometric mean cell volume

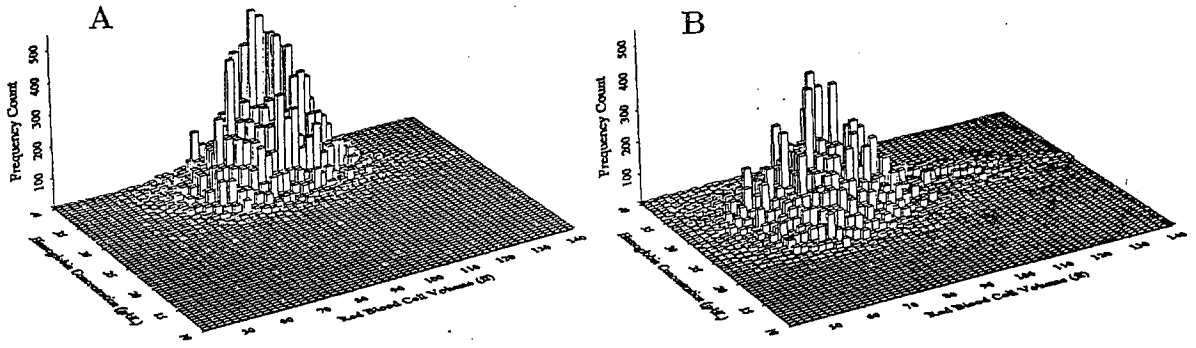

Figure 1: Red blood cell volume and hemoglobin concentration distributions.
A: *Healthy male.* Parameter estimates: mixing proportion = 1.0, geometric mean cell volume = 89 fl, geometric mean cell hemoglobin concentration = 34.4 fl.
B: *Developing iron deficiency anemia.*Parameter estimates: mixing proportion = .58, geometric mean red cell volume = 73.9 fl, geometric mean cell hemoglobin concentration = 28.4 g/dL; mixing proportion = .42, geometric mean red cell volume = 80.9 fl, geometric mean cell hemoglobin concentration = 30.2 g/dL.

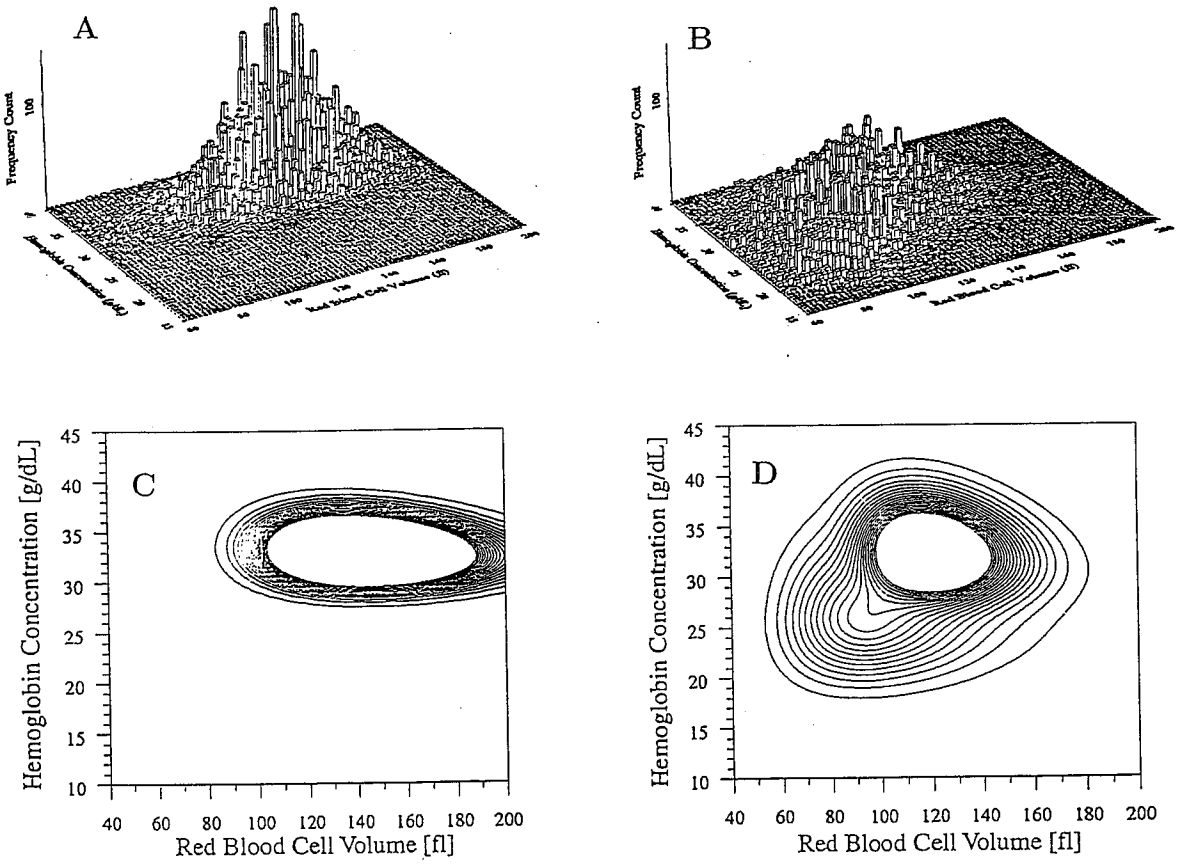

Figure 2: Bivariate distribution and contour plots for red blood cell volume and hemoglobin concentration.
A, C: *Alcoholic Liver Disease.* Parameter estimates: mixing proportion = .95, geometric mean red cell volume = 140 fl, geometric mean cell hemoglobin concentration = 33.0 g/dL; mixing proportion = .05, geometric mean red cell volume = 95.9 fl, geometric mean cell hemoglobin concentration = 32.0 g/dL
B, D: *B_{12} deficiency, folate deficiency, and iron deficiency.* Parameter estimates: mixing proportion = .48, geometric mean red cell volume = 120.7 fl, geometric mean cell hemoglobin concentration = 32.4 g/dL; mixing proportion = .52, geometric mean red cell volume = 94.0 fl, geometric mean cell hemoglobin concentration = 26.0 g/dL.

Table 1: Classification of Healthy Individuals (Controls) and Patient Subgroups

Subgroup	Percent Correct	Number of Cases Classified into Group			Total
		Control	Microcytosis	Macrocytosis	
Control	96.7%	87	1	2	90
Microcytosis	100.0%	0	106	0	106
Macrocytosis	95.0%	1	1	38	40

and hemoglobin concentration of 89 fl and 34.4 g/dL repectively. For comparison, the distribution shown in Figure 1B is from a female with developing iron deficiency anemia. The bivariate distribution contained a hypochromic, microcytic subpopulation of 58% of cells with an estimated geometric mean red cell volume of 73.9 fl and geometric mean hemoglobin concentration of 28.4 g/dL, both below normal. A hypochromic, normocytic subpopulation with estimated 42% of cells had geometric mean red cell volume of 80.9 fl, within the normal range, and geometric mean hemoglobin concentration of 30.2 g/dL.

The distribution and contour plot from a female with alchoholic liver disease with hypochromic, macrocytic anemia are shown in Figure 2A and 2C. The larger subpopulation contains 95% of the cells with an estimated geometric mean cell volume of 140 fl and moderately reduced geometric mean hemoglobin concentration of 33.0 g/dL. A remaining, 5% of the cells had geometric mean volume (95.9 fl) and hemoglobin concentration (32.0 g/dL) consistent with that of controls. There was no correlation between volume and hemoglobin concentration. The estimated correlation coefficient was 0 for both subpopulations (Figure 2C). The bivariate distribution and contour plot (Figure 2B, 2D) from a female patient with B_{12}/folate deficiency represent a hypochromic, macrocytic, subpopulation of 48% of cells with an estimated geometric mean red cell volume of 120.7 fl and geometric mean hemoglobin concentration of 32.4 g/dL and a hypochromic, normocytic, subpopulation containing concentration of 94.0 fl and 26.0 g/dL respectively.

Table 1 gives the number and percent of all patients correctly classified using leave-one-out cross validation. Distributions from three healthy individuals were misclassified as having anemia and one patient with alcoholic liver disease was misclassified as a control. These results are not unexpected, because production of red blood cells is a dynamic process making it difficult to classify distributions falling at the upper or lower limits of the a particular subgroup. The distribution shown in Figure 2B, from a female patient with B_{12}/folate deficiency, was misclassified as having microcytosis. Class posterior probabilities were $< .001$ for classification as control, 1.0 for classification as microcytic, and $< .001$ for classification as macrocytic. Further review of the patient's medical chart revealed that in addition to the low levels of vitamin B_{12} and red cell folate, this patient had reduced serum ferritin values, consistent with iron deficiency anemia. In this instance, the distribution analysis picked up this anomalous case of B_{12} deficiency, folate deficiency, and iron deficiency.

4. DISCUSSION

We have successfully developed techniques to model and classify multivariate distributions from grouped and truncated red blood cell data. Our study is unique in using mixture models in a bilevel fashion, first for description on an individual subject level and then for classification on a group level. On the individual subject level, the range of distributions in healthy individuals with normocytic, normochromic cells is defined, while in patients with anemia, distributions containing subpopulations of cells with microcytic, normocytic, or macrocytic red blood cell volume and hypochromic or normochromic hemoglobin concentration are described. On the group level, parameter estimates for individual mixture models are then used to distinguish between healthy individuals and patients with anemia.

Discrimination between patient subgroups on the basis of the distribution parameters for hemoglobin concentration and red blood cell showed that controls are well separated from other patients with disorders of anemia (Table 1: 98% overall correct classification). In previous studies, classification of patients with thalassemia trait and iron deficiency anemia [12] or vitamin B12/folate deficiency, alcohol excess/liver disease and reticulocytosis [13] utilized printed statistical and graphical output from flow

cytometric blood cell counting instruments. Our study is the first to model the joint distribution of red cell volume and hemoglobin concentration reflecting measurements of individual blood cells. Development of bilevel modeling techniques revealed that within-individual mixtures of red cell subpopulations form between-individual clusters on the basis of disease category.

As described in the Appendix, the bivariate mixture models for each individual can be estimated in a straightforward and computationally efficient manner, using the Expectation-Maximization algorithm. The method can be readily implemented in software on a standard PC workstation, or could equally well be embedded within a flow cytometric blood cell counting instrument. We conclude that for individual subjects, these methods may provide a means for monitoring the response to therapy or for automated screening for disorders of anemia.

APPENDIX

The bilevel model used in this paper consists of two "levels." The lower (individual) level model consists of a two-component lognormal mixture fitted to each of the individual cytograms. Maximum likelihood estimates of the lognormal mixture parameters are obtained for each individual cytogram using the EM procedure as outlined below. Variability among parameters of different individuals is then modeled at the higher (group) level of the hierarchy by a multivariate normal density function for each of the three groups.

Modeling at the Individual Level: EM for Fitting Mixtures to Cytograms

The binned, and in some cases, truncated nature of the cytogram data for each individual requires that the standard EM estimation framework for finite mixtures be somewhat modified. The theory for fitting finite mixture models to such data in the univariate case was developed in full by McLachlan and Jones [8]. Here we present a brief summary of the underlying ideas. The model can be written as:

$$f(x;\Phi) = \sum_{i=1}^{g} \pi_i f_i(x;\theta), \qquad (A.1)$$

where the π_i's are weights for the individual components, the f_i's are the component density functions of the mixture model parametrized by θ, and Φ is the set of all mixture model parameters, $\Phi = \{\pi, \theta\}$. The overall sample space $\mathcal{H}$ is divided into v disjoint subspaces $\mathcal{H}_j$, (bins) of which only the counts on the first r bins are observed, while the counts on last $v-r$ bins are missing. The (observed) likelihood associated with this model (up to irrelevant constant terms) is given by Jones and McLachlan [9]:

$$\ln L = \sum_{j=1}^{r} n_j \ln P_j - n \ln P, \qquad (A.2)$$

where n_j is the count in bin j, n is the total observed count $n = \sum_{j=1}^{r} n_j$, and the Ps represent integrals of the probability density function (PDF) over bins:

$$P_j \equiv P_j(\Phi) = \int_{\mathcal{H}_j} f(x;\Phi)dx, \qquad (A.3)$$

$$P \equiv P(\Phi) = \int_{\mathcal{H}} f(x;\Phi)dx = \sum_{j=1}^{r} P_j. \qquad (A.4)$$

The form of the likelihood function above corresponds to a multinomial distributional assumption on bin occupancy.

In Cadez, Smyth et al. we provide a detailed description of how the EM algorithm can be implemented efficiently in the multidimensional case for binned and truncated data, including the application of the method to mixture modeling of cytograms. The efficiency is achieved by leveraging a variety of computational short-cuts at various stages of the algorithm. For example, for any fixed sample size, a multivariate histogram will be much sparser than any marginal univariate counterpart, in terms of counts per bin (i.e., marginals) and this sparseness can in turn be taken advantage of for the purposes of efficient numerical integration [10].

Group Level Modeling in Parameter Space

The output of the EM mixture modeling is a set of 11 parameters for each individual that describes two-component distributions of red blood cells per individual, consisting of:

- Two component mixing weights (proportions) giving 1 independent parameter as the weights add up to one,
- Two 2-dimensional means representing mean volume and mean hemoglobin concentration of each of the mixture components, yielding 4 additional independent parameters, and
- Two 2×2 covariance matrices describing the "shape" (the distribution around mean) of the cell volume and the cell hemoglobin concentration for each of the mixture components, for

The modeling task is to use these 11-dimensional cytogram parameters to build a 3-component normal mixture model, one component for each of the three groups of interest: controls, macrocytic and microcytic patients. The mixing proportions represent the prior probabilities of belonging to each group. Each mean contains information about the proportion, means and shapes of the cell subpopulations that are likely to be seen in a typical representative of the respective patient group. The 11×11 covariance matrices for each group represent the natural variability among cytograms from patients within the same group. Since the class labels are known a priori, maximum likelihood parameter estimation for each group can be performed in closed form directly. For the results in this paper, the group covariance matrices were assumed to be diagonal. Classification of a new cytogram then consists of a two-step procedure. First, a two-component lognormal mixture model is fit to obtain an 11-dimensional parameter vector as described earlier. The posterior probabilities of group membership are then obtained using Bayes rule and the three 11-dimensional normal densities described previously.

ACKNOWLEDGEMENTS

This work has been supported in part by research grants from the National Institutes of Health (R15-HL48349, R43-HL46037) and a Wellcome Research Travel Grant awarded by the Burroughs Wellcome Fund (CEM). Additional support was provided by a National Science Foundation CAREER award, IRI-9703120 (PS, IVC) and a grant from the Australian Research Council, A10027060 (GJM). We are grateful to Brian Ortner and Dr. Albert Greenbaum for technical assistance. We thank Thomas H. Cavanagh for providing laboratory facilities.

REFERENCES

1. DeMaeyer E. and Adiels-Tegman M. 'The prevalence of anemia in the world', *World Health Statistical Quarterly,* **38**, 302–316 (1985).

2. Oski, F. A. 'Iron deficiency in infancy and childhood', *New England Journal of Medicine,* **329**, 190–193 (1993).

3. Edgerton V. R., Gardner G. W., Ohira Y., Gunawardena K. A., Senewiratne B. 'Iron-deficiency anaemia and its effect on worker productivity and activity patterns.', *British Medical Journal,* **2**, 1546–1549 (1979).

4. Williams, W. J., Beutler, E., Erslev, A. J., and Lichtman, M. A. *Hematology*, fourth edition, McGraw-Hill, New York, 1988.

5. McLaren, C.E., Wagstaff, M., Brittenham, G.M. and Jacobs, A. 'Detection of two component mixtures of lognormal distributions in grouped doubly-truncated data: analysis of red blood cell volume distributions', *Biometrics,* **47**, 607–622 (1991).

6. McLaren, C. E. 'Mixture models in hematology', *Statistical Methods in Medical Research,* **5**, 129–153 (1996).

7. McLaren, C.E., Kambour, E., McLachlan, G.J., Lukaski, H.C., X. Li, Brittenham, G.M., and McLaren, G.D. 'Patient-specific analysis of sequential hematological data by multiple linear regression and mixture distribution modeling', *Statistics in Medicine,* **19**, 83–98 (2000).

8. McLachlan, G.J. and Jones P.N. 'Fitting mixture models to grouped and truncated data via the EM algorithm', *Biometrics,* **44**, 571–578 (1988).

9. Jones, P. N., McLachlan, G. J. 'Maximum Likelihood Estimation from Grouped and Truncated Data with Finite Normal Mixture Models,' *Applied Statistics-Journal of the Royal Statistical Society Series C*, 39(N2):273-282 (1990).

10. Cadez, I. V., Smyth, P., McLachlan, G. J. and McLaren, C. 'Maximum likelihood estimation of mixture densities for binned and truncated multivariate data', *Machine Learning*, (2000), in press.

11. Cadez, I. V., McLaren, C. E., Smyth, P. and McLachlan G. J. 'Hierarchical Models for Screening of Iron Deficiency Anemia', in *Proceedings of the 1999 International Conference on Machine Learning*, I. Bratko and S. Dzeroski (eds.), Los Gatos: CA, Morgan Kaufmann, 77–86,
(1999).

12. Jimenez, C. V., Minchinela, J. and Ros, J.' New indices from the H*2 analyser improve differentiation between heterozygous b or db thalassaemia and iron-deficency anaemia', *Clin Lab Haemat,* **17**, 151–5 (1995).

13. Harkins, L. S., Sirel, J. M., McKay, P. J. and Wylie, R. C., Titterington D. M. and Rowan R, M. 'Discriminant analysis of macrocytic red cells', *Clin Lab haematol,* **16**, 225–234 (1994).