*Appl. Statist.* (1998)
**47**, *Part* 3, *pp.* 393–404

# Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes

Theodore J. Thompson†, Philip J. Smith and James P. Boyle

*Centers for Disease Control and Prevention, Atlanta, USA*

[Received August 1996. Final revision July 1997]

**Summary:** The World Health Organization (WHO) diagnostic criteria for *diabetes mellitus* were determined in part by evidence that in some populations the plasma glucose level 2 h after an oral glucose load is a mixture of two distinct distributions. We present a finite mixture model that allows the two component densities to be generalized linear models and the mixture probability to be a logistic regression model. The model allows us to estimate the prevalence of diabetes and sensitivity and specificity of the diagnostic criteria as a function of covariates and to estimate them in the absence of an external standard. Sensitivity is the probability that a test indicates disease conditionally on disease being present. Specificity is the probability that a test indicates no disease conditionally on no disease being present. We obtained maximum likelihood estimates via the EM algorithm and derived the standard errors from the information matrix and by the bootstrap. In the application to data from the diabetes in Egypt project a two-component mixture model fits well and the two components are interpreted as normal and diabetes. The means and variances are similar to results found in other populations. The minimum misclassification cutpoints decrease with age, are lower in urban areas and are higher in rural areas than the 200 mg dl$^{-1}$ cutpoint recommended by the WHO. These differences are modest and our results generally support the WHO criterion. Our methods allow the direct inclusion of concomitant data whereas past analyses were based on partitioning the data.

*Keywords*: *Diabetes mellitus*; EM algorithm; Finite mixture; Generalized linear model; Sensitivity; Specificity

## 1. Introduction

In 1980, the World Health Organization (WHO) published diagnostic criteria for *diabetes mellitus* (World Health Organization Expert Committee on Diabetes Mellitus, 1980). The criteria for diabetes are either a plasma glucose level greater than or equal to 200 mg dl$^{-1}$ 2 h after an oral glucose load of 75 mg or a fasting plasma glucose level greater than or equal to 140 mg dl$^{-1}$. The WHO cutpoints were chosen, in part, because it had been shown that the 2-h plasma glucose level in some populations can be modelled as a mixture of two distributions (Omar *et al.*, 1994; Rosenthal *et al.*, 1985; Rushforth *et al.*, 1971; Zimmett and Whitehouse, 1978). In these studies, mixture distributions were fitted to subsets of the data defined by sex and age groups. We propose a regression model that can be used to estimate mixture distributions in the presence of concomitant data. This model allows the sensitivity and specificity of the current diagnostic criteria to be estimated in the absence of an external standard.

Finite mixture models were extensively discussed by Everitt and Hand (1981), McLachlan and Basford (1988) and Titterington *et al.* (1985). Bayesian estimation of the parameters of a

†*Address for correspondence*: Division of Diabetes Translation, Mailstop K-10, Centers for Disease Control and Prevention, 4770 Buford Highway NE, Atlanta, GA 30341-3724, USA.
E-mail: tat5@cdc.gov

0035–9254/98/47393

mixture is given in Gilks *et al.* (1996). Using regression models for components in a mixture is a natural extension (e.g. Aitkin and Wilson (1980)). The term generalized linear finite mixture model is due to Jansen (1993), who proposed embedding generalized linear models (GLMs) (McCullagh and Nelder, 1989) into the EM algorithm (Dempster *et al.*, 1977) for fitting mixtures. This approach requires fitting two GLMs, each with $nm$ observations. The number of observations in the original data set is $n$, and the number of distinct distributions (components) in the mixture is $m$. Our approach uses $m+1$ separate GLMs—one for the mixing density and one for each component density (all with $n$ observations). As opposed to the methods of Jansen (1993), our methods allow the component densities to be unrelated. Also the component densities can be of the same type (e.g. normal) and have differing scale parameters. An advantage of Jansen's approach is that the location parameters can be the same across component densities. The ECM algorithm (Meng and Rubin, 1993) allows different scale parameters within Jansen's approach.

We analyse a subset of data from the diabetes in Egypt project. This project was a population-based household survey conducted in Egypt between 1991 and 1994. Results were reported by Herman *et al.* (1995) and Thompson *et al.* (1996). The data analysed here are for all people between the ages of 36 and 65 years who were examined in the clinic portion of the survey. People younger than 36 years of age were excluded because of their low prevalence of diabetes and people older than 65 years of age were excluded to remove the potentially confounding effects of mortality.

We were thus left with 919 people, and we have 919 observations on each of the following variables. The response variable is the 2-h plasma glucose measurement. Glucose measurements 2 h after a 75 g oral glucose load were made by the glucose oxidase method with a Kodak Ektachem DT60 analyser at the Diabetes Institute in Cairo, Egypt. The five covariates are age (36–65 years), sex, residence (urban or rural), level of habitual physical activity (active or sedentary) and obesity (no or yes). Obesity is defined according to the WHO criterion of a body mass index greater than or equal to 30 kg $\mathrm{m}^{-2}$. This binary classification of body mass index was of particular interest to the health professionals and epidemiologists involved with the diabetes in Egypt project. We are interested in the performance of the current WHO criterion of a 2-h plasma glucose measurement greater than or equal to 200 mg $\mathrm{dl}^{-1}$ for the diagnosis of diabetes.

Section 2 of this paper presents the generalized linear finite mixture model, the EM algorithm, covariance matrix estimation, sensitivity and specificity estimates and a goodness-of-fit test. Section 3 applies the model to data from the observational study of diabetes conducted in Egypt, and Section 4 discusses the results.

## 2. Finite mixture models

### 2.1. Generalized linear finite mixture model

In our model we use random variables of two types: a manifest random variable $Y$, which can be directly observed, and a latent discrete random variable $Z$, which is unobservable. As only $Y$ can be observed, inference is based on the marginal density:

$$f(y) = \sum_z h(z) f(y|z)$$

where $h(\cdot)$ is the density function for $Z$. However, we are interested in what can be known about $Z$ after $Y$ is observed. The relevant conditional density is

$$h(z|y) = h(z) f(y|z)/f(y).$$

To specify a model, we need to make assumptions about $h(z)$ and $f(y|z)$.

Assume that each person can be classified into one of two components that denote the presence or absence of disease. Let the Bernoulli random variable $Z$ denote component membership and let $\Pr(Z = 1) = \pi$ (where Pr denotes probability). Let the random variable $Y$ denote a continuous measurement obtained from each person, and let $f(y|Z = 1)$ be denoted by $f_1(y)$ and $f(y|Z = 0)$ by $f_2(y)$. We develop the likelihood equations and their solution via the EM algorithm for $f_1$ and $f_2$ members of the exponential family.

The data consist of a random sample of size $n$. Associated with each observation $y_i$ is a sampling weight or frequency $w_i$ and a row vector of covariates $\mathbf{x}_i$. The $w_i$ is proportional to the probability that observation $i$ is included in the sample, and often $\Sigma\, w_i = n$. An alternative normalization, $\Sigma\, w_i = \Sigma\, w_i^2$, is given by Potthoff *et al.* (1992). In a simple random sample $w_i = 1, \forall i$. Including the weights in the model development is useful even for the case of simple random sampling because it simplifies the bootstrap computations for standard errors as shown below. The log-likelihood for the observed data $y_i$ is

$$l_{\text{obs}} = \sum_{i=1}^{n} w_i \log\{\pi_i f_1(y_i) + (1 - \pi_i) f_2(y_i)\}. \tag{1}$$

We treat the values of $Z$ from our sample $(z_i)$ as missing data and use the EM algorithm for estimation. The complete data log-likelihood for $(z_i, y_i)$ is

$$l_c = \sum_{i=1}^{n} w_i\{z_i \log(\pi_i) + (1 - z_i)\log(1 - \pi_i)\} + \sum_{i=1}^{n} w_i z_i \log\{f_1(y_i)\} + \sum_{i=1}^{n} w_i(1 - z_i)\log\{f_2(y_i)\}. \tag{2}$$

This log-likelihood can be maximized by separately maximizing the three summations.

If we assume that $f_1$ and $f_2$ are members of the exponential family, the EM algorithm together with standard software for GLMs can be used to maximize $l_{\text{obs}}$. Covariates can be included in the densities $h, f_1$ and $f_2$ in the usual way.

### 2.2. Estimation via the EM algorithm

The EM algorithm is implemented as follows. In the M-step the complete data log-likelihood is maximized by replacing $z_i$ with $p_i^{(k)} = E(z_i|y_i)$. This (estimated conditional) probability is calculated in the $k$th iteration of the E-step. Starting values were obtained by setting $p_i^{(0)} = 1$ if $y_i < c$ and $p_i^{(0)} = 0$ otherwise (where $c$ is a specified constant). A series of values for $c$ can be used to avoid problems with convergence to a local maximum. Convergence is obtained when the change in successive values of the observed log-likelihood (equation (1)) is less than $\epsilon$. We set the value of $\epsilon$ to $10^{-5}$.

#### 2.2.1. *M-step*

The following three GLMs are solved. Here we use normal distributions for $f_1$ and $f_2$. In practice any GLM can be used. Different subsets of the covariates $\mathbf{x}_i$ can be included in each of the three GLMs. These are denoted $\mathbf{x}_{0i}$, $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$.

(a) Model 1: Bernoulli distribution for $Z$—use $p_i^{(k-1)}$ as the 'response data', a logistic link function, covariates $\mathbf{x}_{0i}$ and weights $w_i$ to estimate

$$\pi_i^{(k)} = \exp(\mathbf{x}_{0i}\boldsymbol{\alpha}^{(k)})/\{1 + \exp(\mathbf{x}_{0i}\boldsymbol{\alpha}^{(k)})\}.$$

(b) Model 2: normal distribution for $Y$ in component 1—use $y_i$ as the response data, an identity link function, covariates $\mathbf{x}_{1i}$ and weights $w_i p_i^{(k-1)}$ to estimate $\mu_{1i}^{(k)} = \mathbf{x}_{1i}\boldsymbol{\beta}^{(k)}$ and $\sigma_1^{(k)}$. Calculate $f_1^{(k)}(y_i)$ using these estimates.

(c) Model 3: normal distribution for $Y$ in component 2 — use $y_i$ as the response data, an identity link function, covariates $\mathbf{x}_{2i}$ and weights $w_i(1 - p_i^{(k-1)})$ to estimate $\mu_{2i}^{(k)} = \mathbf{x}_{2i}\gamma^{(k)}$ and $\sigma_2^{(k)}$. Calculate $f_2^{(k)}(y_i)$ using these estimates.

At convergence the final estimates are denoted $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_1$, $\hat{\gamma}$, $\hat{\sigma}_2$ and $\hat{\pi}_i$.

### 2.2.2. *E-step*

The conditional probability that observation $i$ comes from component 1 is estimated by

$$p_i^{(k)} = \pi_i^{(k)} f_1^{(k)}(y_i)/\{\pi_i^{(k)} f_1^{(k)}(y_i) + (1 - \pi_i^{(k)}) f_2^{(k)}(y_i)\}.$$

At convergence the final estimates are denoted $\hat{p}_i$, $i = 1, \ldots, n$.

## 2.3. *Covariance matrix*

A disadvantage of maximum likelihood estimation via the EM algorithm is that an estimated covariance matrix of the parameters is not a by-product of the process. Bootstrap methods (Efron, 1979) can be used but they are computationally intensive for this model. A bootstrap replicate of the parameter estimates can be obtained by replacing only $w_i$ and then fitting the model. For example, if all the weights are 1, then sample with replacement from the integers 1, 2, . . ., $n$. Let $w_i^*$ equal the number of times that $i$ is in this sample, and fit the model by changing only $w_i$ to $w_i^*$. This procedure saves rebuilding the $y_i$ and the three design matrices used in the GLMs for each bootstrap replicate.

Louis (1982) proposed a method for estimating the covariance matrices when using the EM algorithm, and McLachlan and Basford (1988) described its application when fitting mixture models. Let $\mathbf{I}(\phi)$ be the matrix defined by

$$\mathbf{I}(\phi) = -\partial^2 l_{\text{obs}}(\phi)/\partial\phi\partial\phi'$$

where $l_{\text{obs}}(\phi)$ denotes the log-likelihood for the observed data as a function of all parameters $\phi = (\alpha', \beta', \sigma_1, \gamma', \sigma_2)'$ in the model. (The primes denote transpose.) Furthermore, let

$$\mathbf{h}(\phi;\, y_i,\, z_i) = \partial l_{\text{c}}(\phi;\, y_i,\, z_i)/\partial\phi$$

where $l_{\text{c}}(\phi;\, y_i,\, z_i)$ is the $i$th element in the summation of equation (2) and the estimate $\hat{\mathbf{h}}_i = \mathbf{h}(\hat{\phi};\, y_i,\, \hat{p}_i)$. The observed information matrix $\mathbf{I}(\hat{\phi})$ is approximated by

$$\mathbf{I}(\hat{\phi}) \approx \sum_{i=1}^{n} w_i^{-1} \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i'$$

when the weights $w_i$ are integer frequencies. Differentiating $l_{\text{c}}(\phi)$ given a logistic mixing density and normal component densities leads to $\hat{\mathbf{h}}_i = (\hat{\mathbf{h}}_{0i}', \hat{\mathbf{h}}_{1i}', \hat{\mathbf{h}}_{2i}')'$ where

$$\hat{\mathbf{h}}_{0i}' = w_i(\hat{p}_i - \hat{\pi}_i)\mathbf{x}_{0i},$$

$$\hat{\mathbf{h}}_{1i}' = \left[\frac{w_i\hat{p}_i(y_i - \mathbf{x}_{1i}\hat{\beta})}{\hat{\sigma}_1^2}\mathbf{x}_{1i},\ \frac{w_i\hat{p}_i\{(y_i - \mathbf{x}_{1i}\hat{\beta})^2 - \hat{\sigma}_1^2\}}{\hat{\sigma}_1^3}\right],$$

$$\hat{\mathbf{h}}_{2i}' = \left[\frac{w_i\hat{p}_i(y_i - \mathbf{x}_{2i}\hat{\gamma})}{\hat{\sigma}_2^2}\mathbf{x}_{2i},\ \frac{w_i\hat{p}_i\{(y_i - \mathbf{x}_{2i}\hat{\gamma})^2 - \hat{\sigma}_2^2\}}{\hat{\sigma}_2^3}\right].$$

Inverting $\mathbf{I}(\hat{\phi})$ yields an estimate of the covariance matrix of $\hat{\phi}$.

### 2.4. Relative sensitivity and specificity

Relative sensitivities and specificities for any given cutpoint $y_0$ are easily calculated from the fitted distributions. We refer to these sensitivity and specificity estimates as *relative* because they are estimated relative to a fitted model and are specific to the population analysed. Sensitivity is the probability that a test indicates disease conditionally on disease being present:

$$\widehat{\mathrm{Se}}(y_0, \mathbf{x}_{2i}) = \int_{y_0}^{\infty} f_2(u; \mathbf{x}_{2i}, \hat{\gamma}, \hat{\sigma}_2)\,\mathrm{d}u. \tag{3}$$

Here we explicitly include covariates $\mathbf{x}_i$ since the densities $f_1$ and $f_2$ are modelled as functions of covariates. Specificity is the probability that a test indicates no disease conditionally on no disease being present:

$$\widehat{\mathrm{Sp}}(y_0, \mathbf{x}_{1i}) = \int_{-\infty}^{y_0} f_1(u; \mathbf{x}_{1i}, \hat{\boldsymbol{\beta}}, \hat{\sigma}_1)\,\mathrm{d}u. \tag{4}$$

When normal distributions are used, standard functions can be used to calculate the sensitivity and specificity. Receiver operating characteristic (ROC) curves (Hanley and McNeil, 1982) may also be obtained from the fitted models. An ROC curve is a plot of Se *versus* $1 - \mathrm{Sp}$ for a range of cutpoints $y_0$.

Methods for the selection of a cutpoint are discussed by McNeil *et al.* (1975). We choose the point that minimizes misclassification. The misclassification rate is defined by

$$\pi\{1 - \mathrm{Se}(\cdot)\} + (1 - \pi)\{1 - \mathrm{Sp}(\cdot)\},$$

and its minimum occurs when $\pi f_1(\cdot) = (1 - \pi) f_2(\cdot)$. Cutpoints that minimize misclassification have also been used by Rosenthal *et al.* (1985).

### 2.5. Goodness of fit

To perform a goodness-of-fit test for a given model, the range of $Y$ is divided into intervals with approximately equal numbers of observations. The probability that an observation is in an interval is calculated separately for each observation by each interval. Within each interval we sum over observations to obtain an expected value for that interval. The usual statistic, $\Sigma_i\,(O_i - E_i)^2/E_i$, is compared with the $\chi^2$-distribution with $n - k - 1$ degrees of freedom, where $i$ indexes the intervals, $O_i$ is the observed value and $E_i$ is the expected value. The number of observations is $n$ and the number of intervals is $k$. Following DeGroot (1975) we choose as many intervals as possible without allowing the expected number in any interval to become small.

## 3. Application

The data analysed include 919 observations on each of six variables. The response variable $y_i$ is the 2-h plasma glucose measurement. The covariates $\mathbf{x}_i$ are age (36–65 years), sex, residence (urban or rural), physical activity (active or sedentary) and obesity (no or yes).

Mixtures of two normal distributions have previously been fitted to the logarithm of 2-h plasma glucose measurements from other populations. Fig. 1 shows histograms of log (2-h plasma glucose) by sex and two age groups (bars) as well as the fitted density for the mixture of two normal distributions (curve). The mixtures appear to fit the data well. A hypothesis test of a single normal *versus* a mixture of two normal distributions was rejected for all four sex-by-age groups (details not presented here).

Insight into the appropriateness of the normal assumption for the logarithm of 2-h plasma glucose can be obtained by using the Box–Cox transformation (Box and Cox, 1964)

$$g(y) = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \ln(y) & \text{if } \lambda = 0. \end{cases}$$

Note that $\lim_{\lambda \to 0}\{(y^\lambda - 1)/\lambda\} = \ln(y)$ and that the derivative of $g$ $(\mathrm{d}g/\mathrm{d}y)$ is $y^{\lambda-1}$. If the probability density function of $g(y)$ is $f(\cdot)$, the probability density function of $y$ is $f\{g(y)\}|\mathrm{d}g/\mathrm{d}y|$. Box and Cox (1964) suggested using the profile likelihood for the largest model to be considered as a guide for choosing $\lambda$ and then using that value of $\lambda$ in subsequent model fitting and selection.

Box–Cox transformations in normal mixture models have been used in other studies. In Gutierrez *et al.* (1995), a mixture model was not needed after transformation. Jansen and Den Nijs (1993) used data of known component membership to estimate the appropriate transformation.

We estimated the optimal transformation within the Box–Cox family of transformations. All covariates were included in all three regression models from Section 2.2. A series of models was fitted with values of $\lambda$ ranging from −1 to 1. The profile likelihood is plotted in

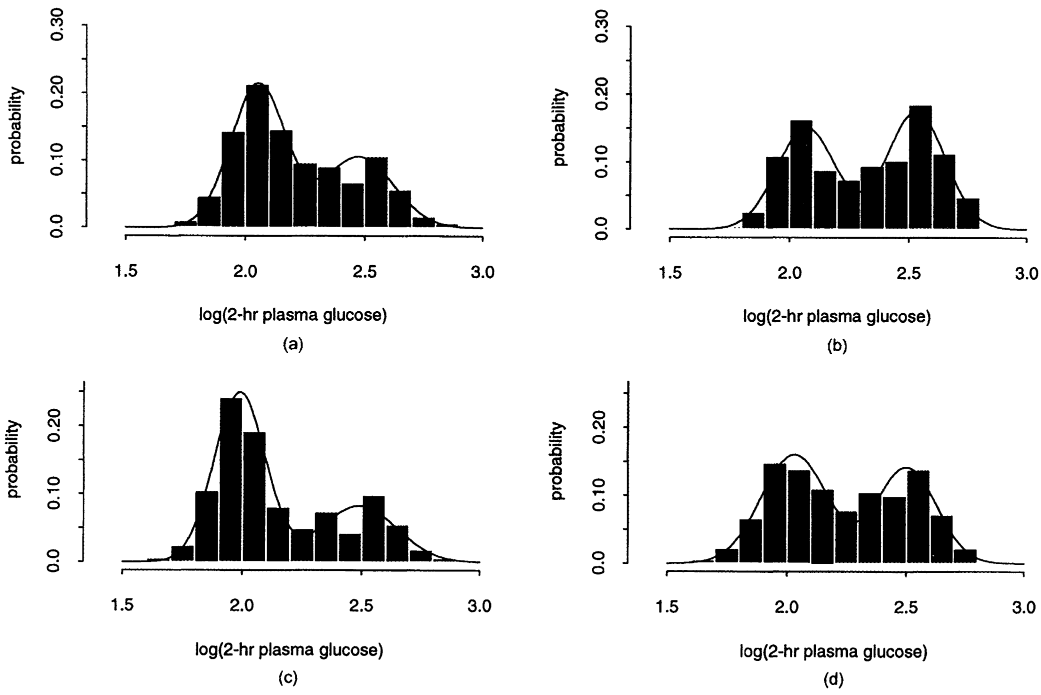

**Fig. 1.** Histograms of log(2-h plasma glucose) by sex and age groups with fitted densities for a mixture of two normal distributions (———): (a) women, 36–50 years old; (b) women, 51–65 years old; (c) men, 36–50 years old; (d) men, 51–65 years old

Fig. 2. The maximum value of the profile likelihood (dotted line) and a 95% confidence interval for the maximum (full line) are shown. The logarithmic transformation ($\lambda = 0$) is close to optimal and well within the 95% confidence interval. We used the $\log_{10}$-transformation for all subsequent analyses.

The basic model included all five covariates in the mixing distribution, in the first-component distribution and in the second-component distribution. There are $(2^5)^3$ or 32768 subsets of this model. Additionally we want to examine all two-way interactions and squared terms in age. Since there were too many possible models to examine all possible subsets, we used a stepwise approach. Starting with the basic model we included interactions and a squared term in age one at a time. We then omitted terms that had no significant interactions, again one at a time. We used the Bayesian information criterion (Schwarz, 1978) for model selection.

Table 1 presents the final model. A goodness-of-fit test based on 75 intervals yielded a $\chi^2$-value of 64.64 on 62 degrees of freedom and $p = 0.38$. The logistic regression model for $\Pr(Z_i = 1) = \pi_i$ includes terms for age, age$^2$ and residence. A comparison of the final model with a model that includes only an intercept in the mixing distribution yielded a likelihood ratio statistic of 105.1 on 3 degrees of freedom. A person is more likely to be in the second component (diabetes) as age increases (an odds ratio of 2.4 at age 50 years *versus* age 40 years and an odds ratio of 1.2 at age 60 years *versus* age 50 years) and is more likely to be in the second component if living in an urban area (an odds ratio of 3.9). The least squares regression model for $\mu_{1i}$ includes terms for residence, obesity and sex nested within physical activity. A comparison of the final model with a model that includes only an intercept in the first component yielded a likelihood ratio statistic of 70.9 on 4 degrees of freedom. The sex by physical activity interaction term was significant but an examination of the other model terms revealed no sex differences among the inactive individuals. Therefore in our final model we collapsed the inactive males and inactive females into a single category. In component one (no diabetes) rural dwellers, obese individuals and physically active females have a higher 2-h

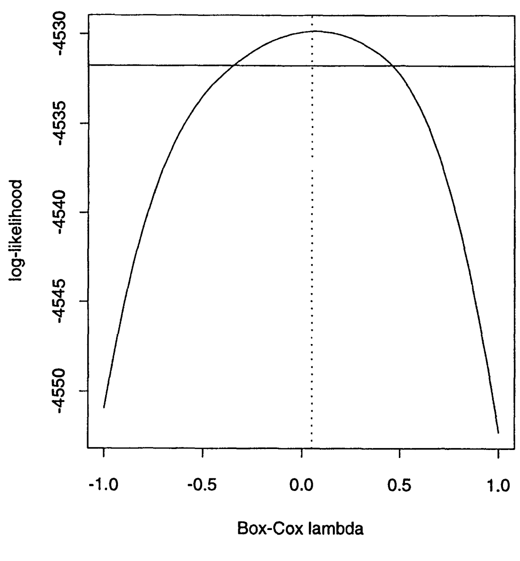

**Fig. 2.** Profile log-likelihood *versus* Box–Cox $\lambda$ with maximum (··········) and 95% confidence interval for the maximum (———)

**Table 1.** Generalized linear finite mixture model

| *Distribution* | *Parameter* | *Estimate* | *Standard error* | *z-value* | *p-value* | *Bootstrap standard error*† | *% change* |
|---|---|---|---|---|---|---|---|
| Mixing | intercept | 1.142 | 0.1939 | 5.89 | | 0.2064 | 6.1 |
| | $s$(age)‡ | −0.511 | 0.0993 | −5.15 | <0.001 | 0.1061 | 6.4 |
| | $s(\text{age})^2$ | 0.370 | 0.1215 | 3.04 | 0.002 | 0.1248 | 2.6 |
| | urban | −1.368 | 0.1914 | −7.15 | <0.001 | 0.1996 | 4.1 |
| First component | intercept | 2.024 | 0.0109 | 185.16 | | 0.0116 | 6.0 |
| | urban | −0.052 | 0.0122 | −4.24 | <0.001 | 0.0131 | 6.9 |
| | obese | 0.071 | 0.0123 | 5.80 | <0.001 | 0.0130 | 5.4 |
| | active female | 0.054 | 0.0149 | 3.66 | <0.001 | 0.0164 | 9.1 |
| | active male | −0.050 | 0.0211 | −2.35 | 0.019 | 0.0162 | −30.2 |
| | standard deviation | 0.117 | 0.0048 | 24.59 | | 0.0058 | 17.2 |
| Second component | intercept | 2.510 | 0.0085 | 296.34 | | 0.0110 | 22.7 |
| | standard deviation | 0.132 | 0.0074 | 17.67 | | 0.0074 | 0.0 |

†Based on 2000 bootstrap replications.
‡$s$(age) = (age−50)/10.

plasma glucose level. Physically active males have a lower glucose level. The least squares regression model for $\mu_{2i}$ includes only an intercept term.

Table 1 also contains bootstrap standard errors and the percentage change between the bootstrap standard errors and the information matrix standard errors. The bootstrap standard errors are generally larger.

The posterior probability of diabetes *versus* 2-h plasma glucose is plotted in Fig. 3 for the 919 observations. The posterior probability of diabetes, $p_i = E(z_i|y_i)$, is calculated in the E-step of the EM algorithm. The WHO cutpoint of 200 mg dl$^{-1}$ is denoted by the dotted line. These estimates of $p_i$ agree with medical knowledge about the discriminating power of 2-h plasma glucose measurements (Rosenthal *et al.*, 1985).

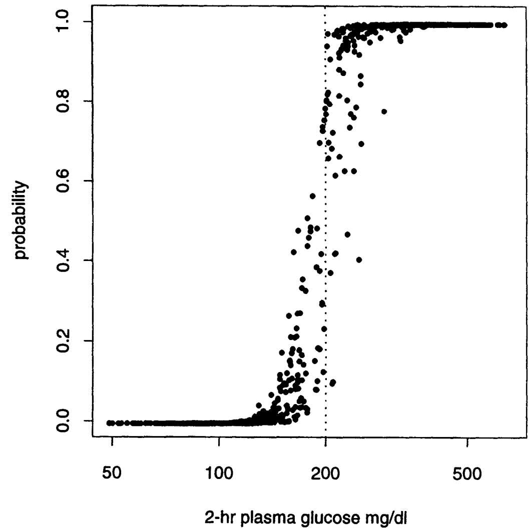

**Fig. 3.** Posterior probability of diabetes based on the final fitted model *versus* 2-h plasma glucose mg dl$^{-1}$

**Table 2.** Cutpoints, sensitivities, specificities and misclassification rates by residence and age†

| *Cutpoint criteria* | *Residence* | *Age (years)* | *Cutpoint (mg $dl^{-1}$)* | *Sensitivity (%)* | *Specificity (%)* | *Misclassification rate (%)* |
|---|---|---|---|---|---|---|
| Minimum misclassification | Rural | 40 | 223 | 89.1 (2.63) | 99.2 (0.80) | 2.0 (0.62) |
| | | 60 | 206 | 93.3 (2.05) | 98.3 (0.66) | 3.1 (0.58) |
| | Urban | 40 | 192 | 95.9 (1.55) | 98.4 (0.66) | 2.5 (0.48) |
| | | 60 | 178 | 97.6 (1.08) | 96.8 (1.06) | 2.7 (0.53) |
| WHO | Rural | 40 | 200 | 94.4 (1.83) | 97.8 (0.80) | 2.6 (0.62) |
| | | 60 | 200 | 94.4 (1.83) | 97.8 (0.80) | 3.1 (0.56) |
| | Urban | 40 | 200 | 94.4 (1.83) | 98.9 (0.50) | 2.6 (0.59) |
| | | 60 | 200 | 94.4 (1.83) | 98.9 (0.50) | 3.7 (1.07) |

†Standard error estimates are given in parentheses.

Table 2 contains estimates of the sensitivity, specificity and misclassification rate at the cut-off points that minimize misclassification and at a cut-off of 200 mg $dl^{-1}$. The estimates were calculated from equations (3) and (4) using the fitted values from Table 1. Bootstrap standard errors for the sensitivity, specificity and misclassification rate are also included. Average values were used for obesity, sex and physical activity. The cutpoints that minimize misclassification are lower in rural areas than in urban areas and these cutpoints decrease with age. In general these cutpoints will decrease as prevalence $1 - \pi$ increases.

## 4. Discussion

We have characterized the distribution of glucose tolerance in the Egyptian population and examined factors that are related to this distribution. Our results are based on a population-based survey conducted in Cairo and surrounding rural villages between 1991 and 1994. This diabetes in Egypt project represents the first time that a comprehensive study of the distribution of glucose tolerance has been conducted in the Egyptian population. The results are similar to findings in other populations (Omar *et al.*, 1994; Rosenthal *et al.*, 1985; Rushforth *et al.*, 1971; Zimmett and Whitehouse, 1978). This consistency of results across populations leads us to believe that our results are reasonable. Since there is no definitive standard for diagnosing diabetes, the current WHO standard was developed via consensus. This study provides additional information for future consensus development.

We presented a model that allows sensitivity and specificity of diagnostic criteria for diabetes to be estimated in the absence of an external standard. This generalized linear finite mixture model also allows covariates to be included in each component density and in the mixing density. This analysis shows that the diagnostic performance of 2-h plasma glucose measurements varies somewhat with age and residence (urban or rural), obesity, physical activity and sex. If we are restricted to selecting a single cutpoint among subpopulations, our results generally support the WHO 2-h criterion.

Age and residence (urban or rural) are related to the mixing (prevalence) distribution. The increase in prevalence of diabetes with age is well known. The large increase in prevalence in urban areas *versus* rural areas is thought to be related to a sedentary life style and obesity. These relationships are somewhat confounded. Obesity and a sedentary life style are risk factors for developing diabetes and they are also a result of insulin treatment for diabetes. Also, the level of physical activity is difficult to measure and obesity is a rather crude binary measurement. However, physical activity and obesity as measured in this survey do not provide a better fitting model for the mixing distribution.

Covariates in the mixing distribution lead to diagnostic criteria that are dependent on those covariates. As presented earlier, choosing a cutpoint that minimizes misclassification requires an estimate of prevalence in the target population. When the prevalence is a function of covariates, adjusting cut-off values for these covariates can potentially improve the performance of the test. Clinicians have long recognized that disease prevalence and other information (e.g. age) affect the choice of a cutpoint used with an imperfect diagnostic test. For example, with rare diseases cutpoints that favour specificity are often used. Here a small decrease in specificity can result in a large number of false positive results. The distributions of many laboratory tests change with characteristics of the individual. A blood urea nitrogen concentration of 25 mg per 100 ml is usually high for a young person but normal for an older person (Martin *et al.*, 1975). Diagnostic tests based on blood urea nitrogen would need to account for an individual's age. Engelgau *et al.* (1995) estimated cutpoints for a capillary blood glucose test that vary by age and post-prandial period.

No covariates are related to the mean of the second (diabetes) component. This supports the idea that 2-h plasma glucose measurements in a diseased population are unrelated to the presence or absence of concomitant factors. Urban *versus* rural residence, obesity, physical activity and sex are related to the mean of the first (non-diabetes) component. Risk factors related to changes in 2-h plasma glucose among non-diseased individuals have not been extensively studied. Obese subjects tend to have higher values and physically active subjects lower values. Other studies have shown these to be risk factors for diabetes. We observed higher values among obese people. The relationship of physical activity was not consistent. Lower values among physically active individuals were only observed for males. Among females physical activity was related to higher 2-h blood glucose measurements. This interaction between sex and physical activity suggests that the physical activity question has a different meaning for males and females or is being interpreted differently by males and females.

Maximum likelihood estimation is straightforward using the EM algorithm and software for GLMs. For example, these models could be programmed in S-PLUS or GLIM. All the calculations for this paper were programmed in GAUSS (Aptech Systems, 1995) and we used an iteratively reweighted least squares algorithm (Aitkin *et al.*, 1989) in the embedded logistic regression model.

We undertook Box–Cox analysis to determine whether log-transformations are appropriate in the current context. Earlier papers (Omar *et al.*, 1994; Rosenthal *et al.*, 1985; Rushforth *et al.*, 1971; Zimmett and Whitehouse, 1978) on mixtures of glucose measurements used log-transformations with no formal justification. Our work shows that the log-transformation is appropriate within the Box–Cox class.

The EM algorithm sometimes requires many iterations before convergence is achieved. During the bootstrap estimation of standard errors, the EM algorithm averaged 39 iterations with a maximum of 106. During each iteration a logistic regression and two least squares problems are solved. However, interactive model fitting is possible with an efficient language (e.g. GAUSS; Aptech Systems (1995)) on a fast personal computer. On a 150-MHz Pentium-based personal computer, the average time required to fit a single model was less than 2 s.

The likelihood surface for mixture models may contain multiple local maxima. McLachlan and Basford (1988) recommended fitting a series of models with differing starting values and using the solution with the largest likelihood. The likelihood surface in the analysis presented here was well behaved. Varying the starting values over a reasonable range always led to the same solution.

## Acknowledgements

The authors thank Dr Michael M. Engelgau, the Joint Editor, the Associate Editor and two referees for helpful comments and suggestions. We also thank William H. Herman of the University of Michigan at Ann Arbor and Mohamed A. Ali and colleagues of the Ministry of Health in Cairo, Egypt, for their efforts during the diabetes in Egypt project.

## References

Aitkin, M., Anderson, D., Francis, B. and Hinde, H. (1989) *Statistical Modelling in GLIM*. Oxford: Clarendon.

Aitkin, M. and Wilson, G. T. (1980) Mixture models, outliers, and the EM algorithm. *Technometrics*, **22**, 325–331.

Aptech Systems (1995) *The GAUSS System Version 3.2.13*. Maple Valley: Aptech Systems.

Box, G. E. P. and Cox, D. R. (1964) The analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

DeGroot, M. H. (1975) *Probability and Statistics*. Reading: Addison-Wesley.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Efron, B. (1979) Computers and the theory of statistics: thinking the unthinkable. *SIAM Rev.*, **21**, 460–480.

Engelgau, M. M., Thompson, T. J., Smith, P. J., Herman, W. H., Aubert, R. E., Gunter, E. W., Wetterhall, S. F., Sous, E. S. and Ali, M. A. (1995) Screening for diabetes mellitus in adults. *Diab. Care*, **18**, 463–466.

Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions*. London: Chapman and Hall.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Guterierrez, R. G., Carroll, R. J., Wang, N., Lee, G.-H. and Taylor, B. H. (1995) Analysis of tomato root initiation using a normal mixture model. *Biometrics*, **51**, 1461–1468.

Hanley, J. A. and McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Herman, W. H., Ali, M. A., Aubert, R. E., Engelgau, M. M., Kenny, S. J., Gunter, E. W., Malarcher, A. M., Brechner, R. J., Wetterhall, S. F., DeStefano, F., Thompson, T. J., Smith, P. J., Badran, A., Sous, E. S., Habib, M., Hegazy, M., abd el Shakour, S. and el Moneim el Behairy, A. (1995) Diabetes mellitus in Egypt — risk factors and prevalence. *Diab. Med.*, **12**, 1126–1131.

Jansen, R. C. (1993) Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics*, **49**, 227–231.

Jansen, R. C. and Den Nijs, A. P. M. (1993) A statistical mixture model for estimating the proportion of unreduced pollen grains in perennial ryegrass (*Lolium perenne* L.) via the size of pollen grains. *Euphytica*, **70**, 205–215.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

Martin, H. F., Gudzinowicz, B. J. and Fanger, H. (1975) *Normal Values in Clinical Chemistry*. New York: Dekker.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

McNeil, B. J., Keeler, E. and Adelstein, S. J. (1975) Primer on certain elements of medical decision making. *New Engl. J. Med.*, **293**, 211–215.

Meng, X. L. and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Omar, M. A. K., Seedat, M. A., Dyer, R. B., Motala, A. A., Knight, L. T. and Becker, P. J. (1994) South African Indians show a high prevalence of NIDDM and bimodality in plasma glucose distribution patterns. *Diab. Care*, **17**, 70–73.

Potthoff, R. F., Woodbury, M. A. and Manton, K. G. (1992) "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *J. Am. Statist. Ass.*, **87**, 383–396.

Rosenthal, M., McMahan, C. A., Stern, M. P., Eifler, C. W., Hazuda, H. P. and Franco, L. J. (1985) Evidence of bimodality of two hour plasma glucose concentrations in Mexican Americans: results from the San Antonio heart study. *J. Chron. Dis.*, **38**, 5–16.

Rushforth, N. B., Bennett, P. H., Steinberg, A. G., Burch, T. A. and Miller, M. (1971) Diabetes in the Pima Indians: evidence of bimodality in glucose tolerance distributions. *Diabetes*, **20**, 756–765.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Thompson, T. J., Engelgau, M. M., Herman, W. H., Ali, M. A., Sous, E. S. and Badran, A. (1996) The onset of NIDDM and its relationship to clinical diagnosis in Egyptian adults. *Diab. Med.*, **13**, 337–340.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

World Health Organization Expert Committee on Diabetes Mellitus (1980) Second report on diabetes mellitus. *WHO Tech. Rep. Ser.*, **28**, 1039–1057.

Zimmett, P. and Whitehouse, S. (1978) Bimodality of fasting and two-hour glucose tolerance distributions in a Micronesian population. *Diabetes*, **27**, 793–800.