Research Report
Department of Statistics
Göteborg University
Sweden

Prediction of work resumption in theory and practice

Anders Persson

Research Report 2002:5
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Dept of Statistics	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.stat.gu.se/stat
P.O. Box 660	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
SE 405 30 Göteborg			
Sweden			

PREDICTION OF WORK RESUMPTION IN THEORY AND PRACTICE

by Anders Persson

Department of Statistics, Göteborg University, Sweden

In Sweden, the number of long-term sick-listed has increased by about 30% per year during the period 1997-2001, and the cost for health insurance is 108 billion SEK in the state budget (2002). Thus, the prediction of work resumption is very important. It is also important to identify the influence of different factors. In this thesis, an approach for prediction of future work resumption is proposed. The suggested method takes into account the dependency structure of the predictors in a flexible way.

In the first paper (1) an approach based on Bayes theorem is proposed for predicting a binary outcome conditionally on the values of a set of discrete predictors. Point- and interval estimators are derived for this probability, and the properties of these estimators are examined in detail by theoretical results and simulations. It is found that the variance of the estimated probability is heavily dependent on the situation. It was also found that a sample size of at least 400 was required to obtain reliable estimates of the prediction probabilities.

The second paper (2) is an application of the suggested approach in paper (1) to predict the outcome of work resumption for men and women with lower back- and neck pain in a Swedish population. In this application, the predictors have a complex dependency structure. Hierarchical cluster analysis has been used to identify independent groups of predictors such that predictors within groups are dependent but independent of predictors in other groups. In a first step, point- and interval estimates of the probability of 'no work resumption' given the value of a set of predictors were calculated from the data set. In a second step, new observations were generated with the same characteristics as those in the first step. Predictive- and relative predictive values as well as proportions of correct classifications were calculated. The predictive values and the proportions of correct predictions ranged from 0.59 to 0.81 and 0.70 to 0.86, respectively.

Papers included in the thesis:

(1) Jonsson, R. and Persson, A. (2002) *Bayes Prediction of Binary Outcomes Based on Correlated Discrete Predictors*. Research Report 2002:3, Department of Statistics, Göteborg University.

(2) Persson, A. (2002) *Prediction of Work Resumption Among Men and Women with Lower Back- and Neck Pain in a Swedish Population.* Research Report 2002:4, Department of Statistics, Göteborg University.

ACKNOWLEDGEMENTS

I would like to acknowledge a special appreciation to my supervisor Ph. D. Robert Jonsson for suggesting the topic and encouraging guidance and support during the work of this thesis. Thanks to Professor Marianne Frisén for valuable support and comments; to Christian Sonesson for constructive comments and suggestions on earlier versions of the papers; to Catrin Eklöf-Olsson and Eva Bang for administrative help, and finally, thanks to all my friends and colleagues at the Department of Statistics, Göteborg University.

Address for correspondence:
Anders Persson, Department of Statistics, Göteborg University, Box 660, SE-405 30 Göteborg, Sweden
Email: Anders.Persson@statistics.gu.se

BAYES PREDICTION OF BINARY OUTCOMES BASED ON CORRELATED DISCRETE PREDICTORS

by Robert Jonsson and Anders Persson

Department of Statistics, Göteborg University, Sweden

Abstract

An approach based on Bayes theorem is proposed for predicting the binary outcomes $X = 0, 1$, given that a vector of predictors $\mathbf{Z}$ has taken the value $\mathbf{z}$. It is assumed that $\mathbf{Z}$ can be decomposed into g independent vectors given $X = 1$ and h independent vectors given $X = 0$. First, point and interval estimators are derived for the target probability $\mathbb{P}(X = 1 \mid \mathbf{z})$. In a second step these estimators are used to predict the outcomes for new subjects chosen from the same population. Sample sizes needed to achieve reliable estimates of the target probability in the first step are suggested, as well as sample sizes needed to get stable estimates of the predictive values in the second step. It is also shown that the effects of ignoring correlations between the predictors can be serious. The results are illustrated on Swedish data of work resumption among long-term sick-listed individuals.

Key words: Conditional independence; Confidence intervals; Interactions; Multinomial probabilities; Prediction; Work resumption.

1 Introduction

In many situations there is a great need for predicting categorical outcomes at the individual level. For example, during recent years there has been an increasing rate of cases with long-term sickness in many countries, and in Sweden the increase has been about 30% per year during the period 1997-2001 (SOU (2002)). This has focused on the need for better individual predictions of future state of health, which in term would facilitate the proper rehabilitating interventions. Commonly used methods for such predictions have been logistic regression (Cox (1970)) or 'computer diagnosis' based on empirical Bayes weights (Afifi and Azen (1979), pp. 306-10). The latter two approaches give identical results, since they only differ in the way in which the predictor variables are represented. With a few exceptions, the two approaches have been used under the assumption that the predictors are independent. The reasons for such an assumption are seldom declared, except for the need for simplification, even if it has been pointed out that the assumption may be unrealistic in most applications (Afifi and Azen (1979), p. 307). The effects of assuming predictors to be independent, when they actually are dependent, upon bias and precision of the estimated parameters and on the prediction error seems to have been ignored.

In this paper we suggest an approach based on Bayes theorem for predicting the two outcomes 'healthy' ($X = 0$) and 'non-healthy' ($X = 1$). The vector of predictors $\mathbf{Z}$ have discrete elements and these are allowed to be dependent in such a way that there are dependency between some predictors and independency between some sets of predictors. Furthermore, the number of independent sets of predictors given $X = 0$ may be different from the corresponding number given $X = 1$. In a first step point and interval estimators are derived for the probability $\mathbb{P}(X = 1 \mid \mathbf{z})$, where $\mathbf{z}$ denotes an outcome of the vector $\mathbf{Z}$. The performance of the estimators are studied in simulations (Section 3 and Section 4). Then, in a second step the estimates are used to predict the outcomes for new subjects being sequentially chosen from the same population (Section 5). The success of the predictions is studied by simulations from which the agreement between predicted and actual outcomes are summarized by the predictive values for the outcomes $X = 0$ and $X = 1$, as well as the proportion of correct predictions. Special attention is devoted to the sample size needed to get reliable estimates of $\mathbb{P}(X = 1 \mid \mathbf{z})$ in the first step, but also to the sample size needed to get stable estimates of the predictive values in the second step. In the simulation study data from a study, called the ISSA-project, will be used (Bergendorff et al. (1997), (2001) and Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset (1997)). In the latter, work resumption among sick-listed men and women with lower back- and neck pain was considered. Here, 5-10 predictors were chosen from more than 200 variables. The extraction of predictors from the original list of variables was made by simply choosing those variables for which a change in the variable value caused the largest change in the empirical probability of work resumption. The variables selection process will not be considered in this paper. Instead attention will be paid to the problem of how to use a given number of predictors in an optimal way. These issues are further considered in (Persson (2002)). The paper finally ends with a discussion in Section 6.

2 Notations and Some Basic Results

Let the binary outcome variable X denote the health state for a given individual, 'non-healthy' ($X = 1$) and 'healthy' ($X = 0$), with probability $p^{(x)} = \mathbb{P}(X = x)$, $x = 0, 1$. Groups of predictors such that elements within groups are dependent and elements in different groups are independent will be called independent groups. In general, it will be assumed that the complete vector of predictors $\mathbf{Z}$ can be decomposed into g independent groups of predictors given $X = 1$, $\mathbf{Z}_1, ..., \mathbf{Z}_g$ and h independent groups given $X = 0$, $\mathbf{Z}_1, ..., \mathbf{Z}_h$. The conditional probabilities are defined as

$$\begin{aligned} \mathbb{P}(\mathbf{Z}_r = \mathbf{z}_r \mid X = x) &= q^{(x)}(\mathbf{z}_r) \text{ and} \\ \mathbb{P}(\mathbf{Z}_s = \mathbf{z}_s \mid X = x) &= q^{(x)}(\mathbf{z}_s), \end{aligned} \tag{1}$$

where $x = 0, 1$, $r = 1, ..., g$ and $s = 1, ..., h$. Thus,

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid X = x) = q^{(x)}(\mathbf{z}) = \left\{ \begin{array}{l} \prod_{r=1}^{g} q^{(1)}(\mathbf{z}_r) \\ \prod_{s=1}^{h} q^{(0)}(\mathbf{z}_s) \end{array} \right. .$$

The observed frequencies corresponding to the outcomes in (1) are denoted by $N^{(x)}(\mathbf{z}_r)$ and $N^{(x)}(\mathbf{z}_s)$, respectively. Obviously, $\sum_{\mathbf{z}} N^{(x)}(\mathbf{z}) = N^{(x)}$, $x = 0, 1$ and $N^{(1)} + N^{(0)} = n$, the fixed total sample size. The above notations are illustrated in Table 1 for the case with two binary predictors.

		$Z_2 \mid X = x$		
		0	1	
$Z_1 \mid X = x$	0	$N^{(x)}(0,0), q^{(x)}(0,0)$	$N^{(x)}(0,1), q^{(x)}(0,1)$	$N_1^{(x)}(0), q_1^{(x)}(0)$
	1	$N^{(x)}(1,0), q^{(x)}(1,0)$	$N^{(x)}(1,1), q^{(x)}(1,1)$	$N_1^{(x)}(1), q_1^{(x)}(1)$
		$N_2^{(x)}(0), q_2^{(x)}(0)$	$N_2^{(x)}(1), q_2^{(x)}(1)$	$N^{(x)}, 1$

Table 1: Cell frequencies and probabilities with two predictor variables, where $x = 0, 1$.

The probability of interest is $\pi = \mathbb{P}(X = 1 \mid \mathbf{z})$, and from Bayes theorem it follows that

$$\pi = \frac{\mathbb{P}(X = 1) \cdot \mathbb{P}(\mathbf{Z} \mid X = 1)}{\sum_x \mathbb{P}(X = x) \cdot \mathbb{P}(\mathbf{Z} \mid X = x)} = \frac{A}{1 + A}, \text{ where } A = \frac{p^{(1)} q^{(1)}(\mathbf{z})}{p^{(0)} q^{(0)}(\mathbf{z})}. \qquad (2)$$

Note that the quantities π and A in (2) are functions of $\mathbf{z}$ although this notation has been suppressed for convenience. Thus, with k binary predictors there are 2^k possible outcomes for π and A.

When all predictors are independent, both conditionally on $X = 1$ and on $X = 0$, then $q^{(x)}(\mathbf{z})$ is a product of the marginal probabilities. For practical reasons it is often a great advantage if conditional independency between predictors, or at least between sets of predictors, can be assumed. This is because empty individual cells are more likely to appear than empty marginal cells, and under independency the probability π can be estimated from marginal frequencies with greater accuracy than from within-cell frequencies. For example, with 11 binary predictors there are $2^{11} = 2048$ individual cells, in contrast to $2 \cdot 11 = 22$ marginal cells. In addition to the case with no independent sets of predictors and the case with independent predictors, there are a variety of cases with partial independency.

The conditional variable $\left(N^{(x)}(\mathbf{z}) \mid N^{(x)} = n^{(x)}\right)$ is obviously multinomially distributed with parameters $n^{(x)}$ and $\mathbf{q}^{(x)}$, where $\mathbf{q}^{(x)}$ is vector of all possible probabilities which have been assigned to $\mathbf{Z}$. Thus, for binary predictors $\mathbf{q}^{(x)} = \left(q^{(x)}(1, ..., 1), ..., q^{(x)}(0, ..., 0)\right)$. The probability generating function (pgf) of $M\left(n^{(x)}, \mathbf{q}^{(x)}\right)$ can be expressed as

$$\mathsf{E}\left[\prod_{z_1 \dots z_k} \left(s^{(x)}_{z_1 \dots z_k}\right)^{N^{(x)}(\mathbf{z})} \mid N^{(x)} = n^{(x)}\right] = \left[\mathbf{s}^{(x)}\left(\mathbf{q}^{(x)}\right)^T\right]^{n^{(x)}}$$

where $\mathbf{s}^{(x)} = \left(s^{(x)}_{1 \dots 1}, ..., s^{(x)}_{z_1 \dots z_k}, ..., s^{(x)}_{0 \dots 0}\right)$ and $\left(\mathbf{q}^{(x)}\right)^T$ is the transpose of $\mathbf{q}^{(x)}$.

Lemma 1 *The vector of all cell frequencies* $\left(N^{(1)}(\mathbf{z}) \vdots N^{(0)}(\mathbf{z})\right)$ *is multinomially distributed with parameters* $\left(n, p^{(1)}\mathbf{q}^{(1)} \vdots\, p^{(0)}\mathbf{q}^{(0)}\right)$.

Proof of Lemma 1.

$$\mathsf{E}\left[\prod_{z_1 \dots z_k} \left(s^{(1)}_{z_1 \dots z_k}\right)^{N^{(1)}(\mathbf{z})} \prod_{z_1 \dots z_k} \left(s^{(0)}_{z_1 \dots z_k}\right)^{N^{(0)}(\mathbf{z})} \mid N^{(1)} = n^{(1)}\right]$$

$$= \mathsf{E}\left[\prod_{z_1 \dots z_k} \left(s^{(1)}_{z_1 \dots z_k}\right)^{N^{(1)}(\mathbf{z})} \mid N^{(1)} = n^{(1)}\right] \cdot \mathsf{E}\left[\prod_{z_1 \dots z_k} \left(s^{(0)}_{z_1 \dots z_k}\right)^{N^{(0)}(\mathbf{z})} \mid N^{(0)} = n - n^{(1)}\right]$$

$$= \left[\mathbf{s}^{(1)}\left(\mathbf{q}^{(1)}\right)^T\right]^{n^{(1)}} \cdot \left[\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^T\right]^{n-n^{(1)}}.$$

Now, $N^{(1)}$ is binomially distributed with parameters n and $p^{(1)}$. Thus, by taking the expectation of the last expression over $N^{(1)}$ we obtain the pgf of

$\left(N^{(1)}(\mathbf{z}) \vdots N^{(0)}(\mathbf{z})\right)$ as

$$\begin{aligned}
&\mathsf{E}\left[\left(\frac{\mathbf{s}^{(1)}\left(\mathbf{q}^{(1)}\right)^{T}}{\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}}\right)^{N^{(1)}}\right]\cdot\left[\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}\right]^{n} \\
&= \left[\left(\frac{\mathbf{s}^{(1)}\left(\mathbf{q}^{(1)}\right)^{T}}{\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}}\right)p^{(1)}+p^{(0)}\right]^{n}\cdot\left[\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}\right]^{n}.
\end{aligned}$$

■

From Lemma 1 it follows that cell frequencies with equal as well as different values of x are negatively correlated. Consider for instance the data in Table 1. Here we obtain,

$$\begin{aligned}
\mathsf{Cov}\left(N^{(1)}(1,1), N^{(1)}(0,0)\right) &= -n\left(p^{(1)}\right)^{2} q^{(1)}(1,1)\, q^{(1)}(0,0) \\
\mathsf{Cov}\left(N^{(1)}(1,1), N^{(0)}(1,1)\right) &= -np^{(1)}\left(1-p^{(1)}\right) q^{(1)}(1,1)\, q^{(0)}(1,1).
\end{aligned}$$

When the predictors are dependent rather than independent, we may, for some combinations of the parameters of $p^{(x)}$ and $q^{(x)}(\mathbf{z})$ obtain extremely different results. To show this we calculate the difference between the probability π in the independent and dependent case. For simplicity and without loss of generality, we consider only the case with two predictors where $Z_1 = 1$ and $Z_2 = 1$. Figure 1 shows the differences for various values of $p^{(1)}/p^{(0)}$ with all possible 2×2 contingency tables with probabilities .05 (.1) .95. The differences are symmetric when $p^{(1)}/p^{(0)} = 1$. Although, it is impossible from Figure 1 to identify the parameter values of $q^{(x)}(\mathbf{z})$, calculations show that the differences tends to zero when the parameter values are similar in both tables i.e. when $q^{(1)}(1,1) \approx q^{(0)}(1,1)$, for all values of $p^{(1)}/p^{(0)}$. The purpose of this illustration is to show that, in fact, it does matter if we assume that the predictors are independent or not.

Expression (2) seems to be the simplest way to express the dependency between π and the q-probabilities, but there are other ways. One is logistic regression.

Consider for example the case with two predictors which are dependent, both given $X = 1$ and $X = 0$. Then,

$$\begin{aligned}
\pi &= \mathbb{P}(X = 1 \mid z_1, z_2) = \frac{A}{1+A}, \text{ where} \\
A &= \frac{p^{(1)}}{p^{(0)}} \left(\frac{q^{(1)}(1,1)}{q^{(0)}(1,1)} \right)^{z_1 z_2} \left(\frac{q^{(1)}(1,0)}{q^{(0)}(1,0)} \right)^{z_1(1-z_2)} \\
&\quad \times \left(\frac{q^{(1)}(0,1)}{q^{(0)}(0,1)} \right)^{(1-z_1)z_2} \left(\frac{q^{(1)}(0,0)}{q^{(0)}(0,0)} \right)^{(1-z_1)(1-z_2)} \\
&= \exp\{\alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2\}, \text{ where} \\
\alpha &= \log \left(\frac{p^{(1)} q^{(1)}(1,1)}{p^{(0)} q^{(0)}(1,1)} \right) \text{ is the intercept,} \\
\beta_1 &= \log \left(\frac{q^{(1)}(1,0)\, q^{(0)}(0,0)}{q^{(0)}(1,0)\, q^{(1)}(0,0)} \right), \beta_2 = \log \left(\frac{q^{(1)}(0,1)\, q^{(0)}(0,0)}{q^{(0)}(0,1)\, q^{(1)}(0,0)} \right) \text{ and} \\
\beta_3 &= \log \left(\frac{q^{(1)}(1,1)\, q^{(0)}(1,0)\, q^{(0)}(0,1)\, q^{(1)}(0,0)}{q^{(0)}(1,1)\, q^{(1)}(1,0)\, q^{(1)}(0,1)\, q^{(0)}(0,0)} \right) \text{ are regression parameters.}
\end{aligned}$$

In a similar way, it can be showed that in the case when the predictors are independent, both conditionally on $X = 1$ and $X = 0$ we obtain

$$\begin{aligned}
A &= \exp\{\alpha + \beta_1 z_1 + \beta_2 z_2\}, \text{ where} \\
\alpha &= \log \left(\frac{p^{(1)}}{p^{(0)}} \prod_{i=1}^{k} q_i^{(x)}(0) \right) \text{ and } \beta_i = \log \left(\frac{q_i^{(1)}(1)\, q_i^{(0)}(0)}{q_i^{(0)}(1)\, q_i^{(1)}(0)} \right)
\end{aligned}$$

for $i = 1, 2, ..., k$, where $q_i^{(x)}(z_i)$ denotes the marginal probabilities. With k dependent predictors there will be $2^k - 1$ β-coefficients, and this way of representing the q-probabilities will be extremely extensive. Notice also that omitting the interactions between the predictors in the logistic model is equivalent to assuming that the latter are independent.

Another approach for parametrization is the use of Bayes weights. Again, assume for simplicity that we have two predictors Z_1 and Z_1, then we may rewrite A in (2) as

$$\begin{aligned} & \exp\left\{\log\left(\frac{p^{(1)}}{p^{(0)}}\right)+\log\left(\frac{q^{(1)}(\mathbf{z})}{q^{(0)}(\mathbf{z})}\right)\right\} \\ = & \exp\left\{\log\left(\frac{p^{(1)}}{p^{(0)}}\right)+\log\left(\frac{q^{(1)}(z_1)}{q^{(0)}(z_1)}\right)+\log\left(\frac{q^{(1)}(z_2)}{q^{(0)}(z_2)}\right)\right\}, \end{aligned}$$

where, $\log\left(q^{(1)}(z_i)/q^{(0)}(z_i)\right)$ are called Bayes weights (Afifi and Azen (1979), p. 306-10).

3 Point Estimation of π

The Maximum Likelihood (ML) estimator of the target probability in (2) is obtained as

$$\hat{\pi}=\frac{\hat{A}}{\hat{A}+1}, \text{ where } \hat{A}=\frac{\left(N^{(0)}\right)^{h-1}\prod_{r=1}^{g}N^{(1)}(\mathbf{z}_r)}{\left(N^{(1)}\right)^{g-1}\prod_{s=1}^{h}N^{(0)}(\mathbf{z}_s)}. \tag{3}$$

Some simple examples of (3) are:

$$(i)\ \hat{A}=\frac{N^{(1)}(\mathbf{z})}{N^{(0)}(\mathbf{z})},\ (ii)\ \hat{A}=\left(\frac{N^{(0)}}{N^{(1)}}\right)^{k-1}\prod_{i=1}^{k}\frac{N_i^{(1)}(z_i)}{N_i^{(0)}(z_i)},\ (iii)\ \hat{A}=\frac{N^{(0)}N_{13}^{(1)}(\mathbf{z})}{N_{12}^{(0)}(\mathbf{z})\,N_3^{(0)}(z_3)}(?!).$$

In (i) no sets of predictors are independent, and in (ii) all predictors are independent. In (iii), Z_1, Z_2 and Z_3 are dependent when $X=1$, while (Z_1, Z_2) and Z_3 are two independent groups of predictors when $X=0$.

The fact that (3) is the ML estimator is a direct consequence of Lemma 1. According to the latter, $N^{(x)}(\mathbf{z})/n$ and $N^{(x)}/n$ are the ML estimators of $p^{(x)}q^{(x)}(\mathbf{z})$ and $p^{(x)}$, respectively, so $N^{(x)}(\mathbf{z})/N^{(x)}$ is the ML estimator of $q^{(x)}(\mathbf{z})$ and from this the result in (3) follows.

Below some properties of the estimator in (3) are studied, and some expressions for the estimated variance are given. Results will be derived separately for the case when all predictors are dependent and for the more general case when g groups of predictors are dependent given $X = 1$ and h groups are dependent given $X = 0$. The reason for the separation of the two cases is that various degrees of approximations are used for deriving the results.

Case I. *No sets of predictors are independent*

The estimator in (3) is now obtained from the special case (i) above and an expression for the variance of the latter is given by

$$\mathsf{Var}\left[\hat{\pi}\right] = \frac{\pi\left(1-\pi\right)}{n}\left(\frac{1}{\pi'} + \frac{\left(1-\pi'\right)}{n\left(\pi'\right)^2}\right) = \frac{\pi\left(1-\pi\right)}{n} \cdot C, \text{ say,} \tag{4}$$

where $\pi' = p^{(1)}q^{(1)}\left(\mathbf{z}\right) + p^{(0)}q^{(0)}\left(\mathbf{z}\right)$. An estimator of the variance (4) is obtained from

$$\widehat{\mathsf{Var}}\left[\hat{\pi}\right] = \frac{\hat{\pi}\left(1-\hat{\pi}\right)}{(n-1)}\left(\frac{1}{\hat{\pi}'} + \frac{\left(1-\hat{\pi}'\right)}{n\left(\hat{\pi}'\right)^2}\right) = \frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n} \cdot \hat{C}, \text{ say,} \tag{5}$$

where $\hat{\pi}' = n^{-1}\left(N^{(1)}\left(\mathbf{z}\right) + N^{(0)}\left(\mathbf{z}\right)\right)$.

In order to motivate these expressions, notice that according to Lemma 1 and the results (A1) and (A2) in the Appendix, it follows that, for a fixed value of $\mathbf{z}$, $N' = N^{(1)}\left(\mathbf{z}\right) + N^{(0)}\left(\mathbf{z}\right)$ is binomially distributed with parameters n and π' and also that $\left(N^{(1)}\left(\mathbf{z}\right)|N'\right)$ is binomially distributed with parameters N' and π. Thus, we obtain the expectation

$$\mathsf{E}\left[\hat{\pi}\right] = \underset{N'}{\mathsf{E}}\left[\mathsf{E}\left(\hat{\pi} \mid N'\right)\right] = \underset{N'}{\mathsf{E}}\left[\frac{N'\pi}{N'}\right] = \pi,$$

so $\hat{\pi}$ is unbiased. The variance is (Rao (1973), p. 97)

$$\mathsf{Var}\left[\hat{\pi}\right] = \underset{N'}{\mathsf{E}}\left[\mathsf{Var}\left(\hat{\pi} \mid N'\right)\right] + \underset{N'}{\mathsf{Var}}\left[\mathsf{E}\left(\hat{\pi} \mid N'\right)\right]$$

$$= \underset{N'}{\mathsf{E}}\left[\frac{N'\pi(1-\pi)}{(N')^2}\right] + \underset{N'}{\mathsf{Var}}[\pi] = \pi(1-\pi)\underset{N'}{\mathsf{E}}\left[(N')^{-1}\right] + 0. \qquad (6)$$

Since there is a non-zero probability $\left[(1-\pi')^n\right]$ that N' takes the value 0, one should re-define the estimator of π either by adding 1 in the denominator or by conditioning on $N' > 0$. This would however make the estimator is unnecessary complicated in the large sample situation which is considered here. Instead a Taylor series expansion will be used. From Appendix (A4) it follows that

$$\underset{N'}{\mathsf{E}}\left[(N')^{-1}\right] \approx \frac{1}{n\pi'} + \frac{(1-\pi')}{(n\pi')^2}. \qquad (7)$$

By inserting the approximate expectation (7) into (6) we obtain the variance in (4). The estimated variance in (5) is obtained by simply replacing the parameters π and π' by their obvious estimators. By using $n-1$ in the denominator rather than n, a slight improvement of the closeness to the true variance is obtained.

The expression for the variance of $\hat{\pi}$ in (4) agreed well with the true variance determined from simulations. However, there were some deviations depending on the sample size n and the parameters $q^{(x)}(\mathbf{z})$. The best agreement was obtained with a uniform distribution of the q-probabilities. A simulation study with four cells as in Table 1, showed that with a uniform distribution, the absolute relative difference was below 1% even for a relatively small sample size $n = 50$, and declined rapidly for larger values of n. The agreement became worse when one of the cell probabilities was close to 1. For example, with the parameter setting $q^{(x)}(1,1) = 0.93$, $q^{(x)}(1,0) = 0.02 = q^{(x)}(0,1)$, $q^{(x)}(0,0) = 0.03$, $x = 0,1$, the absolute relative difference was large as 60% for $n = 50$. In the latter case one has to choose $n = 400$ to keep the absolute relative difference below 5% and to choose $n = 800$ in order to keep it below 0.5%. It was also found that similar conclusions could be drawn about the average performance of the estimated variance in (5) as for (4).

Even though the last example is a rather extreme one, it illustrates that some caution is needed when (4) and (5) are used in situations where the cell probabilities are close to 0 or 1.

By means of (4) it is possible to study analytically how the variance of $\hat{\pi}$ depends on the parameters $p^{(1)}$, $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$. When $p^{(1)} = \frac{1}{2}$ the variance is a symmetric function of $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$ which decreases as the latter of the two quantities increase, as can be seen in Figure 2. For $p^{(1)} \neq \frac{1}{2}$ the behavior of the variance is more complicated. When $p^{(1)} < \frac{1}{2}$ the variance decreases with increasing $q^{(0)}(\mathbf{z})$, but now the variance has a local maximum at some $q^{(1)}(\mathbf{z}) > 0$ (Figure 3). The value of $q^{(1)}(\mathbf{z})$ which gives this maximum will increase as $p^{(1)}$ tends to zero. When $p^{(1)} > \frac{1}{2}$ the same pattern is observed, but with $q^{(1)}(\mathbf{z})$ interchanged by $q^{(0)}(\mathbf{z})$ (Figure 4).

Case II. *g sets of predictors are independent given X=1 and h sets of predictors are independent given X=0*

An expression for the variance of $\hat{\pi}$ is given by

$$\mathsf{Var}\left[\hat{\pi}\right] = \frac{\pi^2(1-\pi)^2}{n}\left\{n\left[\prod_{r=1}^{g}\left(1+\frac{1-q^{(1)}(\mathbf{z}_r)}{np^{(1)}q^{(1)}(\mathbf{z}_r)}\right)\right.\right.$$

$$\left.\left.+\prod_{s=1}^{h}\left(1+\frac{1-q^{(0)}(\mathbf{z}_s)}{np^{(0)}q^{(0)}(\mathbf{z}_s)}\right)-2\right]+\frac{1}{p^{(1)}p^{(0)}}\right\} = \frac{\pi^2(1-\pi)^2}{n}\cdot D, \text{ say.} \quad (8)$$

An estimator of $\mathsf{Var}\left[\hat{\pi}\right]$ is

$$\begin{aligned}\widehat{\mathsf{Var}}\left[\hat{\pi}\right] &= \frac{\hat{\pi}^2(1-\hat{\pi})^2}{(n-1)}\cdot\hat{D}, \text{ where} \quad (9)\\ \hat{D} &= n\left[\prod_{r=1}^{g}\left(1+\frac{1}{N^{(1)}(\mathbf{z}_r)}-\frac{1}{N^{(1)}}\right)\right.\\ &\quad\left.+\prod_{s=1}^{h}\left(1+\frac{1}{N^{(0)}(\mathbf{z}_s)}-\frac{1}{N^{(0)}}\right)-2\right]+\frac{n^2}{N^{(1)}N^{(0)}}\end{aligned}$$

In contrast to Case I, the denominator of $\hat{\pi}$ now consists of a sum of products of multinomial variables and the exact distribution of this is very complicated. Instead all derivations will be based on Taylor approximations.

From Appendix (A4) it follows that

$$\begin{aligned}\mathsf{Var}\left[\hat{\pi}\right] &\approx \frac{\mathsf{Var}\left[\hat{A}\right]}{\left(\mathsf{E}\left[\hat{A}\right]+1\right)^4}, \text{ where} \\ \mathsf{Var}\left[\hat{A}\right] &= \underset{N^{(1)}}{\mathsf{E}}\left[\mathsf{Var}\left(\hat{A} \mid N^{(1)}\right)\right] + \underset{N^{(1)}}{\mathsf{Var}}\left[\mathsf{E}\left(\hat{A} \mid N^{(1)}\right)\right].\end{aligned} \tag{10}$$

$$\text{Now, } \mathsf{Var}\left(\hat{A} \mid N^{(1)}\right) = \frac{\left(N^{(0)}\right)^{2(h-1)}}{\left(N^{(1)}\right)^{2(g-1)}} \cdot \mathsf{Var}\left(\frac{\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right)}{\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right)} \mid N^{(1)}\right),$$

where $\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right)$ and $\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right)$ are two independent products conditionally on $N^{(1)}$. These products consist of independent variables, which are distributed $M\left(N^{(1)}, q^{(1)}\left(\mathbf{z}_r\right)\right)$ and $M\left(N^{(0)}, q^{(0)}\left(\mathbf{z}_s\right)\right)$, respectively. From Appendix (A3) it follows that, for fixed values of $\mathbf{z}_r$ and $\mathbf{z}_s$,

$$\begin{aligned}\mathsf{E}\left(\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right) \mid N^{(1)}\right) &= \left(N^{(1)}\right)^{g} \prod_{r=1}^{g} q^{(1)}\left(\mathbf{z}_r\right), \text{ and} \\ \mathsf{E}\left(\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right) \mid N^{(0)}\right) &= \left(N^{(0)}\right)^{h} \prod_{s=1}^{h} q^{(0)}\left(\mathbf{z}_s\right), \text{ while}\end{aligned}$$

$$\begin{aligned}&\mathsf{Var}\left(\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right) \mid N^{(1)}\right) \\ =\ &\left(N^{(1)}\right)^{2g}\left(\prod_{r=1}^{g} q^{(1)}\left(\mathbf{z}_r\right)\right)^2\left\{\prod_{r=1}^{g}\left(1+\frac{1-q^{(1)}\left(\mathbf{z}_r\right)}{N^{(1)}q^{(1)}\left(\mathbf{z}_r\right)}\right)-1\right\}, \text{ and} \\ &\mathsf{Var}\left(\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right) \mid N^{(1)}\right) \\ =\ &\left(N^{(0)}\right)^{2h}\left(\prod_{s=1}^{h} q^{(0)}\left(\mathbf{z}_s\right)\right)^2\left\{\prod_{s=1}^{h}\left(1+\frac{1-q^{(0)}\left(\mathbf{z}_s\right)}{N^{(0)}q^{(0)}\left(\mathbf{z}_s\right)}\right)-1\right\}.\end{aligned}$$

By using the Taylor expansion in Appendix (A4) it is seen that the variance of any ratio of independent variables X and Y can be written

$$\mathsf{Var}\left(\frac{X}{Y}\right) \approx \left(\frac{\mathsf{E}\left(X\right)}{\mathsf{E}\left(Y\right)}\right)^2\left(\frac{\mathsf{Var}\left(X\right)}{\left[\mathsf{E}\left(X\right)\right]^2}+\frac{\mathsf{Var}\left(Y\right)}{\left[\mathsf{E}\left(Y\right)\right]^2}\right).$$

From the last results and by taking the approximate expectation over $N^{(1)}$ it finally follows that

$$\underset{N^{(1)}}{\mathsf{E}}\left[\mathsf{Var}\left(\hat{A}\mid N^{(1)}\right)\right]\approx A^2\left[\prod_{r=1}^{g}\left(1+\frac{1-q^{(1)}\left(\mathbf{z}_r\right)}{np^{(1)}q^{(1)}\left(\mathbf{z}_r\right)}\right)+\prod_{s=1}^{h}\left(1+\frac{1-q^{(0)}\left(\mathbf{z}_s\right)}{np^{(0)}q^{(0)}\left(\mathbf{z}_s\right)}\right)\right].$$

In a similar way it can be shown that

$$\underset{N^{(1)}}{\mathsf{E}}\left[\hat{A}\mid N^{(1)}\right]\approx\frac{N^{(1)}\prod_{r=1}^{g}q^{(1)}\left(\mathbf{z}_r\right)}{N^{(0)}\prod_{s=1}^{h}q^{(0)}\left(\mathbf{z}_s\right)},$$

and by again using the Taylor approximation in Appendix (A3) one gets

$$\underset{N^{(1)}}{\mathsf{Var}}\left[\mathsf{E}\left(\hat{A}\mid N^{(1)}\right)\right]\approx\frac{A^2}{n}\frac{1}{p^{(1)}p^{(0)}}.$$

The expression for $\mathsf{Var}\left[\hat{\pi}\right]$ in (8) is finally obtained from (10) and by using the fact that $A^2/\left(A+1\right)^4=\pi^2\left(1-\pi\right)^2$.

The estimator of the variance in (9) is simply obtained by inserting obvious estimators for parameters.

When $g=1=h$, the expression in (8) should reduces to (4). However, in this case it is easily shown that (8) can be written as

$$\mathsf{Var}\left[\hat{\pi}\right]=\frac{\pi\left(1-\pi\right)}{n}\frac{1}{\pi'}.$$

Thus, the two expressions in (4) and (8) are the same if

$$\frac{\pi\left(1-\pi\right)}{n^2}\cdot\frac{\left(1-\pi'\right)}{\left(\pi'\right)^2}\approx 0.$$

The agreement between the expressions for the variance of $\hat{\pi}$ in (8), the estimated variance in (9), and the true variance was determined from 100,000 simulations. In this case the comparison is complicated by the fact that there are many q-probabilities involved, and therefore we only consider the case with two independent sets of mutually dependent predictors $\mathbf{Z}_1=\left(Z_1,Z_2\right)$ and $\mathbf{Z}_2=\left(Z_3,Z_4\right)$, both given $X = 1$ and $X = 0$. By varying the parameters $p^{(1)}$, $q_{12}^{(x)}(z_1, z_2)$ and $q_{34}^{(x)}(z_3, z_4)$, $x = 0, 1$, it was found that the absolute difference between the variance of $\hat{\pi}$ in the simulations and the variance given by (8) and (9) with a few exceptions were below .001 for $n \geq 200$. In no case the difference was larger than .0003 for $n \geq 400$. In the sequel we choose $n = 400$ and study how the variance of $\hat{\pi}$ in (8) depends on the magnitude of the q-probabilities and also on the number of independent sets of predictors

Figures 5-12 illustrate how the variance simultaneously depends on $q_{12}^{(1)}(z_1, z_2)$ and $q_{34}^{(1)}(z_3, z_4)$ for some values of $p^{(1)}$, $q_{12}^{(0)}(z_1, z_2)$ and $q_{34}^{(0)}(z_3, z_4)$. All variances are considered for a fixed set of (Z_1, Z_2, Z_3, Z_4), e.g. $(1, 1, 0, 1)$. Therefore, the z-arguments have been omitted in the legends to the figures. In Figure 5 it is seen that the variance is a symmetric function of its arguments when $p^{(1)} = \frac{1}{2}$ and $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot)$. For $p^{(1)} < \frac{1}{2}$ (see Figures 6-12), the pattern is more complex and in this case one can identify a saddle-point. The level of the latter increases as $q_{12}^{(0)}(z_1, z_2) = q_{34}^{(0)}(z_3, z_4)$ tends to zero, while at the same time the saddle becomes tighter. For $p^{(1)} > \frac{1}{2}$ this saddle-point pattern vanishes and the variance increases as $q_{12}^{(0)}(z_1, z_2)$ and $q_{34}^{(0)}(z_3, z_4)$ tends to zero (not shown in the figures).

To study how the variance of $\hat{\pi}$ depends on the number of independent sets of predictors some simplifications have to be made. Put $g = h$, so there is an equal number of sub-groups of independent predictors both given $X = 1$ and given $X = 0$, and assume that all $q^{(1)}(\mathbf{z}) = q^{(1)}$ and $q^{(0)}(\mathbf{z}) = q^{(0)}$ while $p^{(1)} = \frac{1}{2}$. Then Figure 14 shows that the variance of $\hat{\pi}$ increases with increasing g as far as $q^{(1)} = q^{(0)}$, and that the increase is larger for small $q's$. When $q^{(1)} \neq q^{(0)}$ there is a different pattern. For large differences between the $q's$, the variance declines with increasing value of g, but for smaller differences the variance has a local maximum before it starts to decline. These findings suggest that much can be gained if it is possible to find (1) many predictors with the property that (2) the q-probabilities $q^{(1)}(\mathbf{z})$ differ much from $q^{(0)}(\mathbf{z})$. On the other hand, failure to identify predictors with different q-probabilities, or including such predictors for some reasons, will increase the variance of $\hat{\pi}$.

4 Interval Estimation of π

When the estimated value of π is used for predicting the state of an individual, it is customary to make the predictions '$X = 1$' if $\pi > \frac{1}{2}$ and '$X = 0$' if $\pi < \frac{1}{2}$ if the costs of misclassification are unknown. Such rigid classification rules may be useful if one wants to evaluate the prediction ability of certain predictors, but for practical purpose they can be risky. The predicted outcome of an individual sometimes calls for an intervention, by for instance offer the individual medical rehabilitation programs. Wrong predictions may then be very expensive. If the costs of misclassification are known, the rigid rule above can be replaced by generalized Bayes classification rules, which minimize the expected cost of misclassification (Afifi and Azen (1979), p. 292). However, the costs are seldom known, or may be hard to quantify. In such cases it may be wise to compute a confidence interval (CI) for π. CI's that are clearly outside $\frac{1}{2}$, can be considered to indicate that the corresponding predictions are more likely than CI's that cover $\frac{1}{2}$. In this section we consider various ways to construct a CI for π. As in the preceding section, two cases will be treated separately.

Case I. *No sets of predictors are independent*

We will compare the expected length and actual coverage probability of five different CI's. Let T d. as. $N(0,1)$ denote that a statistic T asymptotically has a standard normal distribution. The various CI's are derived from the following properties, where the same notations are used as in Section 3.

$$(i)\ \frac{\hat{\pi}-\pi}{\{\mathsf{Var}\,[\hat{\pi}]\}^{1/2}}\ \text{d. as. } N(0,1), \quad (ii)\ \frac{\hat{\pi}-\pi}{\left\{\frac{\pi(1-\pi)}{n}\cdot\hat{C}\right\}^{1/2}}\ \text{d. as. } N(0,1),$$

$$(iii)\ \frac{\hat{\pi}-\pi}{\left\{\frac{\pi(1-\pi)}{N'}\right\}^{1/2}}\ \text{d. as. } N(0,1),\ (iv)\ \left(N^{(1)}(\mathbf{z})\,|N'\right)\ \text{d. } B(N',\pi)\ \text{and}$$

$$(v)\ \frac{\log\left(\hat{A}\right)-\log(A)}{\left\{\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right]\right\}^{1/2}}\ \text{d. as. } N(0,1).$$

Here the statistics in (iii) and (iv) are conditional and based on the particular outcome $N' = N^{(1)}(\mathbf{z}) + N^{(0)}(\mathbf{z})$, while $\log\left(\hat{A}\right)$ is an estimator of $\log(A)$ to be considered below. Let Z be the $100\,(1-\alpha/2)\,\%$ percentile of the standard normal distribution, and let $F\,(n_1, n_2)$ denote the $100\,(1-\alpha/2)\,\%$ percentile of the F-distribution with n_1 and n_2 degrees of freedom. Then the CI's derived from $(i)-(iv)$ are $\hat{\pi}_L < \pi < \hat{\pi}_U$, where $\hat{\pi}_L$ and $\hat{\pi}_U$ are obtained from:

$$(i)\quad \hat{\pi} \pm Z \cdot \left\{\mathsf{Var}\left[\hat{\pi}\right]\right\}^{1/2}$$

$$(ii)\quad \frac{2\hat{\pi}+\frac{Z^2\hat{C}}{n} \pm \left\{\left(2\hat{\pi}+\frac{Z^2\hat{C}}{n}\right)^2 - 4\hat{\pi}^2\left(1+\frac{Z^2\hat{C}}{n}\right)\right\}^{1/2}}{2\left(1+\frac{Z^2\hat{C}}{n}\right)}$$

$$(iii)\quad \frac{2\hat{\pi}+\frac{Z^2}{N'} \pm \left\{\left(2\hat{\pi}+\frac{Z^2}{N'}\right)^2 - 4\hat{\pi}^2\left(1+\frac{Z^2}{N'}\right)\right\}^{1/2}}{2\left(1+\frac{Z^2}{N'}\right)} \qquad (11)$$

(iv)

$$\begin{aligned}
\hat{\pi}_L &= \frac{N^{(1)}(\mathbf{z})}{\left(N^{(0)}(\mathbf{z})+1\right) F\left[2\left(N^{(0)}(\mathbf{z})+1\right), 2N^{(1)}(\mathbf{z})\right]}, \\
\hat{\pi}_U &= \frac{\left(N^{(1)}(\mathbf{z})+1\right) F\left[2\left(N^{(1)}(\mathbf{z})+1\right), 2N^{(0)}(\mathbf{z})\right]}{N^{(0)}(\mathbf{z}) + \left(N^{(1)}(\mathbf{z})+1\right) F\left[2\left(N^{(1)}(\mathbf{z})+1\right), 2N^{(0)}(\mathbf{z})\right]}
\end{aligned}$$

$$(v)\quad \frac{\exp\left\{\log\left(\hat{A}\right) \pm 1.96\left\{\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right]\right\}^{1/2}\right\}}{1+\exp\left\{\log\left(\hat{A}\right) \pm 1.96\left\{\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right]\right\}^{1/2}\right\}}, \text{ where}$$

$$\begin{aligned}
\log\left(\hat{A}\right) &= \log\left(N^{(1)}(\mathbf{z})\right) - \log\left(N^{(0)}(\mathbf{z})\right) \text{ and} \\
\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right] &= \frac{1}{N^{(1)}(\mathbf{z})} + \frac{1}{N^{(0)}(\mathbf{z})}.
\end{aligned}$$

The expressions $(11):(i)-(iv)$ follows from well known results (Casella and Berger (1990), p. 444-49). $(11):(v)$ follows from very rough approximations (see Appendix (A4)) $\mathsf{E}\left[\log\left(\hat{A}\right)\right] \approx \log(A)$, and

$$\mathsf{Var}\left[\log\left(\hat{A}\right)\right] \approx \frac{\mathsf{Var}\left[N^{(1)}(\mathbf{z})\right]}{\left\{\mathsf{E}\left[N^{(1)}(\mathbf{z})\right]\right\}^2} + \frac{\mathsf{Var}\left[N^{(0)}(\mathbf{z})\right]}{\left\{\mathsf{E}\left[N^{(0)}(\mathbf{z})\right]\right\}^2} - 2\frac{\mathsf{Cov}\left[N^{(1)}(\mathbf{z}), N^{(0)}(\mathbf{z})\right]}{\mathsf{E}\left[N^{(1)}(\mathbf{z})\right]\mathsf{E}\left[N^{(0)}(\mathbf{z})\right]},$$

where

$$\begin{aligned}
\mathsf{E}\left[N^{(x)}(\mathbf{z})\right] &= np^{(x)}q^{(x)}(\mathbf{z}) \\
\mathsf{Var}\left[N^{(x)}(\mathbf{z})\right] &= np^{(x)}q^{(x)}(\mathbf{z})\left(1-p^{(x)}q^{(x)}(\mathbf{z})\right),\ x=0,1, \text{ and} \\
\mathsf{Cov}\left[N^{(1)}(\mathbf{z}),N^{(0)}(\mathbf{z})\right] &= -np^{(1)}p^{(0)}q^{(1)}(\mathbf{z})\,q^{(0)}(\mathbf{z}).
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathsf{Var}\left[\log\left(\hat{A}\right)\right] &\approx \frac{1}{n}\left(\frac{1}{p^{(1)}q^{(1)}(\mathbf{z})}+\frac{1}{p^{(0)}q^{(0)}(\mathbf{z})}\right), \text{ and hence} \\
\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right] &\approx \frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}.
\end{aligned}$$

The simple expression in $(11):(v)$ is worth a comment. $\log\left(\hat{A}\right)$ is in fact a poor estimator of $\log(A)$. By instead using the alternative estimator

$$\log\left(\hat{A}\right)+\frac{1}{2}\left(\frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}\right),$$

which follows by considering terms of the order n^{-1} in the Taylor expansion of $\mathsf{E}\left[\log\left(\hat{A}\right)\right]$, both bias and variance can be reduced substantially. The estimated variance of this alternative estimator is

$$\begin{aligned}
&\frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}-\frac{1}{\left[N^{(1)}(\mathbf{z})\right]^2}-\frac{1}{\left[N^{(0)}(\mathbf{z})\right]^2} \\
&+\frac{1}{4}\left(\frac{1}{\left[N^{(1)}(\mathbf{z})\right]^3}+\frac{1}{\left[N^{(0)}(\mathbf{z})\right]^3}\right)-\frac{1}{4n}\left(\frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}\right)^2.
\end{aligned}$$

To illustrate the difference between the two estimators of $\log(A)$, consider the case when there are 2 dependent predictors Z_1 and Z_2, given $X=1$ and given $X=0$, and with the parameter setting

$$\begin{aligned} q_{12}^{(1)}(1,1) &= .24, q_{12}^{(1)}(1,0) = .38, q_{12}^{(1)}(0,1) = .11, q_{12}^{(1)}(0,0) = .27, \\ q_{12}^{(0)}(1,1) &= .71, q_{12}^{(0)}(1,0) = .25, q_{12}^{(0)}(0,1) = .02, q_{12}^{(0)}(0,0) = .02. \end{aligned}$$

A simulation study using the relatively large sample size of $n = 400$, showed that the alternative estimator had a relative bias which was more than 50% smaller than the original estimator. The variance was reduced by 35% and the expression above for the estimated variance of the alternative estimator was very close to the actual variance. However, when the alternative estimator was used for making CI's, the distribution of the pivotal statistic for (v) was slightly skew, and for this reason the coverage rate of 95% was not maintained. The actual coverage rate could in fact drop down to 91%. This illustrates that a CI based on a crude estimator may perform better than a CI based on a more sophisticated estimator.

The performance of the CI's in $(11) : (i) - (v)$ was found to depend on the q-probabilities. As for the expressions (4) and (5) in Section 3, the worst case was obtained when one of the cell probabilities are close to 1. This is illustrated in Table 2, where the 5 CI's are compared regarding expected length and coverage probability. First of all one may notice that none of the CI's keeps the stipulated level of 95% if the sample size, n, is 100 or less. For $n = 200$ the 95%-level is only maintained by $(11) : (ii)$ and possibly by $(11) : (iii)$. However, the expected lengths of the latter are too large to be accepted. When the q-probabilities tend to be more uniformly distributed, the probability that the 95% level is maintained increases, also for smaller samples. The overall conclusion is that $(11) : (ii)$ performs best, even if the CI's may be somewhat conservative. When n is large the computational simple expression in $(11) : (v)$ may be an alternative. $(11) : (i)$ should be avoided. The latter CI's did not even maintain the 95% level in the most favorable case with uniformly distributed q-probabilities and $n = 1600$.

Case II. *g predictors are independent given X=1 and h predictors are independent given X=0*

Now the CI's are derived from the following properties, where the same notations are used as in Section 3 for Case I:

$$(i)\ \frac{\hat{\pi}-\pi}{\left\{\frac{\pi^2(1-\pi)^2}{n}\cdot\hat{D}\right\}^{1/2}}\ \text{d. as. } N(0,1),\ (ii)\ \frac{\log\left(\hat{A}\right)-\log\left(A\right)}{\left\{\hat{D}/n\right\}^{1/2}}\ \text{d. as. } N(0,1).$$

Due to the complexity of the statistic A in this case, we do not consider any conditional statistics, as in Case I. The CI's of π derived from (i) and (ii) above now are $\hat{\pi}_L < \pi < \hat{\pi}_U$, where $\hat{\pi}_L$ and $\hat{\pi}_U$ are the solutions of

$$(i)\ \frac{\left(z\sqrt{\hat{D}/n}\pm 1\right)\mp\sqrt{\left(z\sqrt{\hat{D}/n}\pm 1\right)^2\mp 4z\hat{\pi}\sqrt{\hat{D}/n}}}{2z\sqrt{\hat{D}/n}} \tag{12}$$

$$(ii)\ \frac{\exp\left\{\log\left(\hat{A}\right)\pm 1.96\left\{\hat{D}/n\right\}^{1/2}\right\}}{1+\exp\left\{\log\left(\hat{A}\right)\pm 1.96\left\{\hat{D}/n\right\}^{1/2}\right\}}$$

In (i) the upper part of the two signs ± and ∓ refers to $\hat{\pi}_L$ and the lower part to $\hat{\pi}_U$. In (ii) the upper part of ± refers to $\hat{\pi}_U$ and the lower part to $\hat{\pi}_L$.

(i) follows from the following arguments. Put $f(\pi) = (\hat{\pi}-\pi)/\left(\pi-\pi^2\right)$. Then the statement $-z < (\hat{\pi}-\pi)/\sqrt{\mathrm{Var}\left[\hat{\pi}\right]} < z$ is equivalent to $-z\sqrt{D/n} < f(\pi) < z\sqrt{D/n}$, where the meaning of D is clear from (8). Here $f(\pi)$ is a monotonously decreasing function of $\pi \in (0,1)$ for all $\hat{\pi} \in (0,1)$ with the inverse

$$\pi = \frac{f+1-\sqrt{(f+1)^2-4f\hat{\pi}}}{2f},$$

which gives the CI in (i).

Now, $\log\left(\hat{A}\right)$ can be written $\log(\hat{\pi})-\log(1-\hat{\pi})$, and by using (8) together with Appendix (A4) one gets

$$\mathsf{E}\left[\log\left(\hat{A}\right)\right] \approx \log(A) + \left(\pi - \frac{1}{2}\right) \cdot \frac{D}{n} \text{ and } \mathsf{Var}\left[\log\left(\hat{A}\right)\right] \approx \frac{D}{n},$$

which motivates the use of the statistic in *(ii)*. The expression for the CI in (12) follows easily by noticing that

$$c_L < \log(A) < c_U \text{ implies that } \frac{\exp\{c_L\}}{1+\exp\{c_L\}} < \pi < \frac{\exp\{c_U\}}{1+\exp\{c_U\}}.$$

When $\hat{D}$ in (12) is used for constructing a confidence interval for π, $N^{(1)}(\mathbf{z}_r)$ and $N^{(0)}(\mathbf{z}_s)$ in (12) should be replaced by $N^{(1)}(\mathbf{z}_r) + 1$ and $N^{(0)}(\mathbf{z}_s) + 1$, respectively. This will make the confidence interval less conservative.

Tables 3 and 4 show expected lengths and coverage probabilities for the two CI's in (12), the latter being determined from simulations. The differences between the two are very small. (12) : *(i)* tends to give somewhat shorter CI's, but (12) : *(ii)* tends to give CI's which agree better with the stipulated level of 95%. Again we point out that, although $\log\left(\hat{A}\right)$ is a poor estimator of $\log(A)$, CI's constructed from $\log\left(\hat{A}\right)$ perform well.

5 Prediction

In this section we consider the possibility to predict the outcomes $X = 1$ and $X = 0$ based on $\hat{\pi}$, the estimates of π. The outcome $X = 1$ will be predicted whenever $\hat{\pi} > \frac{1}{2}$ and otherwise the outcome $X = 0$ will be predicted. This rather strict classification rule is chosen merely for simplicity. In practical work it would perhaps be better to use a less rigid classification rule and take the CI's for π into consideration. The predictions will be performed in a two-step approach, where in the first step π is estimated from a sample of a certain population, and then in a second step this estimate is used to predict the outcomes for new subjects being chosen from the same population. If the predicted outcome is denoted by XP, the success of the predictions will be measured by the predictive values $\mathbb{P}(X = 1 \mid XP = 1)$ and $\mathbb{P}(X = 0 \mid XP = 0)$, and the probability of a correct prediction $\mathbb{P}(Correct)$ (see Ch. 3 in Campbell and Machin (1990)). Of special interest will be to study how the predicting ability depends on the sample size, which is used in the first step to estimate π, and also to determine the sample size, which is needed in the second step for reaching stable estimates of the measures of predicting ability. Attention will also be paid to study how miss specification of the dependency structure of the predictors may affect the predicting ability.

5.1 A Simulation Example

In this section we consider the ability to predict work resumption for long-termed sick-listed subjects. The sample considered here is a part of a larger sample within the ISSA-study that has previously been described in detail (Bergendorff et al. (1997), (2001) and Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset (1997)), and consisted of 545 full-time working employed men sick-listed for at least 28 days because of a lower back pain diagnosis. After 28 days the values on the following predictor variables were obtained: (1) Age, (2) Complete rehabilitation plan, (3) Comorbidity, (4) Working ability, (5) Sick-listing in family, (6) Suitable working tasks, (7) Ethnicity, (8) Heavy lifts. Here, Comorbidity means that the subjects has other diseases than lower back pain. Working ability was subjectively assessed on a scale ranking from 1 (low) to 10 (high). Suitable working tasks means that the employer was willing to adjust the working tasks in agreement with the subject's state of health. In a previous study, these variables were found to be the most important ones for predicting work resumption among men with lower back pain (Bergendorff et al. (2001)).

The outcomes to predict at 90 days are $X = 1$, if there is no work resumption and $X = 0$ otherwise. The predictor variables were dichotomized in the following way. Age $= Z_1 = 1$, if age > 30 years and 0 otherwise, Complete rehabilitation plan $(Z_2) = 1$, if yes and 0 otherwise, Comorbidity $(Z_3) = 1$, if yes and 0 otherwise, Working ability $(Z_4) = 1$, if scale value < 5 and 0 otherwise, Sick-listening in family $(Z_5) = 1$, if yes and 0 otherwise, Suitable working tasks $(Z_6) = 1$, if no and 0 otherwise, Ethnicity $(Z_7) = 1$, if Swedish and 0 otherwise and Heavy lift $(Z_8) = 1$, if yes and 0 otherwise.

Notice that all binary predictors have been defined in such a way that the outcome 1 of a predictor favors the outcome $X = 1$. The reasons for dichotomizing the variables Age and Working ability have given previously (Bergendorff et al. (2001)). Although the variable Age has been found to be continuously negatively related to the probability of work resumption in other studies (Jonsson (2001)), this was not the case in the present study where the selected subjects differed from the test of the population in several aspects. E.g. all were full-time working employed.

In this example the first task is to estimate

$$\pi = \mathbb{P}(X = 1 \mid \mathbf{z}) = \frac{p^{(1)}q^{(1)}(\mathbf{z})}{p^{(1)}q^{(1)}(\mathbf{z}) + p^{(0)}q^{(0)}(\mathbf{z})}.$$

A hierarchical cluster analysis (Anderberg (1973) and Jobson (1992)) suggested the following independent sets of vectors

$$\begin{aligned}(\mathbf{Z} \mid X = 1) &= \{(Z_1, Z_2, Z_3 \mid X = 1), (Z_4, Z_5 \mid X = 1), (Z_6, Z_7, Z_8 \mid X = 1)\} \\ (\mathbf{Z} \mid X = 0) &= \{(Z_1, Z_8 \mid X = 0), (Z_3, Z_4, Z_6 \mid X = 0), (Z_2, Z_5, Z_7 \mid X = 0)\}\end{aligned}$$

Thus, e.g. Z_1 (age) and Z_3 (comorbidity) were correlated among those who did not return to work after 90 days, but uncorrelated among those who returned to work. For a more detailed description of the dependency structures the reader is referred to the paper by Persson (2002). The corresponding q-probabilities were

$$\begin{aligned}q^{(1)}(\mathbf{z}) &= q^{(1)}(z_1, z_2, z_3) \cdot q^{(1)}(z_4, z_5) \cdot q^{(1)}(z_6, z_7, z_8) \\ q^{(0)}(\mathbf{z}) &= q^{(0)}(z_1, z_8) \cdot q^{(0)}(z_3, z_4, z_6) \cdot q^{(0)}(z_2, z_5, z_7)\end{aligned}$$

where,

z_1, z_2, z_3	$q^{(1)}(z_1, z_2, z_3)$	z_4, z_5	$q^{(1)}(z_4, z_5)$	z_6, z_7, z_8	$q^{(1)}(z_6, z_7, z_8)$
111	.09	11	.02	111	.22
110	.02	10	.17	110	.02
101	.50	01	.21	101	.06
011	.01	00	.60	011	.35
100	.27	–	–	100	.01
010	.01	–	–	010	.01
001	.07	–	–	001	.23
000	.03	–	–	000	.10

z_1, z_8	$q^{(0)}(z_1, z_8)$	z_3, z_4, z_6	$q^{(0)}(z_3, z_4, z_6)$	z_2, z_5, z_7	$q^{(0)}(z_2, z_5, z_7)$
11	.64	111	.01	111	.02
10	.24	110	.02	110	.02
01	.11	101	.01	101	.01
00	.01	011	.03	011	.05
–	–	100	.01	100	.01
–	–	010	.24	010	.30
–	–	001	.03	001	.05
–	–	000	.65	000	.54

These q-probabilities were estimated from the data set, and will be used as fixed probabilities for generating samples in the simulation study. The prevalence $p^{(1)}$ was 0.54. This figure was also taken from the empirical study.

The various outcomes $(z_1, ..., z_8)$ give rise to 256 values of the estimated posterior probability π. The 5 smallest and largest of these are

$z_1z_2z_3z_4z_5z_6z_7z_8$	$\mathbb{P}(X = 1 \mid \mathbf{z})$	$z_1z_2z_3z_4z_5z_6z_7z_8$	$\mathbb{P}(X = 1 \mid \mathbf{z})$
11001010	.0156	01010000	.9852
11011010	.0316	01110000	.9852
10000010	.0391	01000100	.9860
11000010	.0432	01000110	.9860
10001010	.0786	01000111	.9860

Here one may notice that $z_1 = 1$ (age > 30 years) in all cases giving the smallest probability, while $z_1 = 0$ in all cases giving the largest probabilities.

The simulation experiment was performed in the following way: First, one sample was selected, each being based on the sample sizes $n = 25, 50, 100, 200, ..., 1000$, and from each sample π was estimated. The latter quantity was then used to predict the outcome at 90 days for new subjects being selected from the same population. The number of new sampled subjects was $m = 1000, ..., 100000$, and for each of these, the outcome $X = 1$ was predicted $(XP = 1)$ if $\hat{\pi} > \frac{1}{2}$, and the outcome $X = 0$ was predicted $(XP = 0)$ if $\hat{\pi} < \frac{1}{2}$. The predicted outcomes were then compared with the actual outcomes, and the predictive values were computed as well as the proportion of correct predictions. Here it was found that the predictive values had stabilized already at $m = 1000$.

Figure 14 shows how the predictive values depend on the sample size n in the first sample. It is seen that the predictive values starts to stabilize when n is larger than 400 and that this stabilization process goes faster for $(XP = 1)$ than for $(XP = 0)$. The final values were 0.74 for $(XP = 1)$, 0.73 for $(XP = 0)$ and 0.73 for $\mathbb{P}(correct)$. The similarity between the latter values is merely a coincidence.

6 Discussion

When predicting the future state of health based on estimated probabilities, the choice of good predictors is of major importance, like in all areas of prediction. If very little is known about which variables that will serve as good predictors, a first step may be to perform preliminary study where as many variables as possible are included as candidates. This was made in the ISSA-study mentioned in Section 1 and 5.1. Here, 5-10 variables were chosen as predictors among a total of more than 200 variables. In this paper we have considered the situation where a first sample is taken in order to estimate π and where the prediction ability is evaluated in a second sample from the same population. Then the questions arise of how to extract the predictors from a larger list of candidates, how many to use and how to identify the dependency structure between them, if necessary. The dependency structure can be created by hierarchical cluster methods (Anderberg (1973) and Jobson (1992)). Simulations show that the procedure works very well with dichotomous variables. Since a correct specification of independent clusters has been showed to be of such great importance this issue should be further investigated.

Throughout the paper it has been assumed that the dependency structures between sets of predictors are correctly specified. This is a matter of crucial importance, since by assuming sets of predictors to be conditionally independent when they in fact are dependent may have serious effects on bias and variance of the estimator of π. An illustrative example is the following one with two predictors. Let the cell probabilities in Table 1 be $q_{12}^{(1)}(1,1) = 0.10$, $q_{12}^{(1)}(1,0) = 0.40 = q_{12}^{(1)}(0,1)$, $q_{12}^{(0)}(1,1) = 0.20$, $q_{12}^{(0)}(1,0) = 0.10 = q_{12}^{(0)}(0,1)$, so that the correlation between Z_1 and Z_2 is -0.60 given $X = 1$ and $+0.52$ given $X = 0$. From (2) it follows that the target probability to estimate when $(z_1, z_2) = (1,1)$ is $\pi = 0.33$, and according to (4) $\mathsf{Var}[\hat{\pi}] = 0.0148$ when $n = 100$. On the other hand, by assuming independency between Z_1 and Z_2 the target probability becomes $\pi = 0.74$, while the variance of the estimator is 0.0067 when $n = 100$. Thus, both bias and variance will in this case differ with about 120%. This was just a counter example, but in practice the effects of ignoring correlations between the predictors can be serious and give rise to large differences between the estimated $\pi's$ (see the discussion in Persson (2002)).

The results in Section 3 support the idea to include as many predictors as possible in the model, provided that the difference between the q-probabilities $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$ is large. When the latter difference is small, it may result in a local increase in the variance of $\hat{\pi}$ (see Figure 14). This argues against using predictors in the model with only slight differences between the q-probabilities. For $p^{(1)} = \frac{1}{2}$ and when both $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$ are small, the variance of $\hat{\pi}$ in (4) will be large, as shown in Figure 2. When there are two independent groups of predictors and $p^{(1)} = \frac{1}{2}$, Figure 5 suggests that the variance of $\hat{\pi}$ will be large if both $q_{12}^{(1)}(z_1, z_2)$ and $q_{34}^{(1)}(z_3, z_4)$ are small. These results should apply to the example in Section 5.1 where $p^{(1)}$ was close to $\frac{1}{2}$. Notice that many of the q-probabilities were small. For $p^{(1)} < \frac{1}{2}$ there is a different pattern. Now, Figures 6-12 suggests that the variance will be large when there is a large difference between the $q^{(1)}$-probabilities.

There are also questions about sample sizes needed to get reliable estimates of model parameters and of predictive values. The variance of $\hat{\pi}$ can be reduced by increasing the sample size, but due to the complicated dependencies on the parameters of the expression for the variance, it is not easy to give clear-cut recommendations for the choice of a proper sample size. The smallest sample size needed to reach an acceptable level of the variances of $\hat{\pi}$, for making reliable CI statements and also for getting reliable values of the predictive values was $n = 400$. The latter may be smaller when the q-probabilities are relatively large, but $n = 400$ may be recommended as a safe rule of thumb. Even with samples of 400 it is seen from Tables 2-4 that the lengths of the CI's can be somewhat large, and that sample sizes above 1000 would be needed in order to get CI's with reasonable lengths.

Although all results of the paper apply to predictors with an arbitrary number of outcomes, we have only been concerned with dichotomized predictors in the example of Section 5.1, and this needs an explanation. The reasons for only using binary predictors were that almost all of the variable values were subjectively assessed on an ordinal scale (exceptions were Age and Income), and that more or less pronounced threshold values could either be detected on probability plots (e.g. Working ability on a 10-point scale), or determined after consulting experts in the field (e.g. Complete rehabilitation plan on a 5-point scale). It was supposed that dichotomized predictors would behave more robustly than the original ordinal variables when predictions were made for new subjects. It may be argued that information is lost by the dichotomization. However, in the present study it was felt that this loss of information could be neglected. For instance, the variable 'Complete rehabilitation plan' got the maximal value 5 if the document was signed by the insured, but 4 if the same document was not signed. Here it seemed to be more relevant to know whether such a document existed or not. A further reason for dichotomizing is to reduce the possibility of getting zero cell frequencies. When there are enough many possible outcomes for a predictor it will be inevitable that this will occur. The problem with zero frequencies and missing values are further considered in Persson (2002).

Acknowledgments

The authors would like to thank Christian Sonesson for helpful comments and valuable suggestions on earlier versions of the manuscript.

References

[1] Afifi, A.A. and Azen, S.P. (1979) *Statistical Analysis - A Computer Oriented Approach* (2nd ed.). New York: Academic Press.

[2] Anderberg, M.R. (1973) *Cluster Analysis for Applications.* New York: Academic Press.

[3] Bergendorff, S., Hansson, E., Hansson, T., Palmer, E., Westin, M. and Zetterberg, C. (1997) (In Swedish) *Projektbeskrivning och undersökningsgrupp.* Rygg och Nacke 1. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[4] Bergendorff, S., Hansson, E., Hansson, T. and Jonsson, R. (2001) (In Swedish) *Vad kan förutsäga utfallet av en sjukskrivning?* Rygg och Nacke 8. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[5] Campbell, M.J. and Machin, D. (1990) *Medical Statistics.* New York: Wiley.

[6] Casella, G. and Berger, R.L. (1990) *Statistical Inference.* Belmont California: Duxbury Press.

[7] Cox, D.R. (1970) *Analysis of Binary Data.* London: Chapman and Hall.

[8] Jobson, J.D. (1992) *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods.* New York: Springer-Verlag.

[9] Jonsson, R. (2001) (In Swedish) *Faktorer som är väsentliga vid arbetslivsinriktad rehabilitering samt deras prognosvärde.* Seminar Paper 2001:4. Department of Statistics, Göteborg University.

[10] Kotz, S. and Johnson, N.L. (1985) In *Encyclopedia of Statistical Sciences,* Vol 8. New York: Wiley.

[11] Persson, A. (2002) *Prediction of Work Resumption Among Men and Women with Lower Back- and Neck Pain in a Swedish Population.* Research Report 2002:4. Department of Statistics, Göteborg University.

[12] Rao, C.R. (1973) *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley.

[13] Riksförsäkringsverket and Sahlgrenska Universitetssjukhuset (1997) (In Swedish) *Enkäter till undersökningsgruppen och försäkringskassan.* Rygg och Nacke 2. Stockholm: Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset.

[14] SOU (2002) (In Swedish) *Handlingsplan för ökad hälsa i arbetslivet.* Statens Offentliga Utredningar, 2002:2. Stockholm: Fritzes.

Appendix

Some results for multinomial distributions.
Let $(X_1^{(1)},...,X_k^{(1)},X_1^{(0)},...,X_k^{(0)})$ be a random vector with a multinomial distribution denoted by $M(n,p^{(1)}q_1^{(1)},...,p^{(1)}q_k^{(1)},p^{(0)}q_1^{(0)},...,p^{(0)}q_k^{(0)})$, where $\sum_{i=1}^k q_i^{(1)} = 1 = \sum_{i=1}^k q_i^{(0)}$ and $p^{(1)}+p^{(0)}=1$. A binomial distribution with parameters n and p is denoted by $B(n,p)$.

From the probability generating function (pgf) it is easily verified that

$$X_i^{(1)}+X_i^{(0)} \text{ is distributed } B(n,p^{(1)}q_i^{(1)}+p^{(0)}q_i^{(0)}),\ i=1,...,k. \tag{A1}$$

Direct calculation yields that

$$(X_i^{(1)} \mid X_i^{(1)}+X_i^{(0)}=x) \text{ is distributed } B\left(x,\frac{p^{(1)}q_i^{(1)}}{p^{(1)}q_i^{(1)}+p^{(0)}q_i^{(0)}}\right),\ i=1,...,k. \tag{A2}$$

Let $N(\mathbf{z}_r)$, $r=1,...,g$, be independent vectors each being distributed $M(n,q(\mathbf{z}_r))$. For fixed $\mathbf{z}_r$, $r=1,...,g$, one may put $N_r=N(\mathbf{z}_r)$ and $q_r=q(\mathbf{z}_r)$. Then

$$\mathsf{Var}\left(\prod_{r=1}^g N_r\right)=n^{2g}\left(\prod_{r=1}^g q_r\right)^2\left\{\prod_{r=1}^g\left(1+\frac{1-q_r}{nq_r}\right)-1\right\} \tag{A3}$$

(A3) follows easily by repeated use of the expressions,

$$\begin{aligned}\mathsf{Var}(N_1) &= nq_1(1-q_1),\\ \mathsf{Var}(N_1N_2) &= \mathsf{Var}(N_1)\,\mathsf{Var}(N_2)+\mathsf{Var}(N_1)\,[\mathsf{E}(N_2)]^2+[\mathsf{E}(N_1)]^2\,\mathsf{Var}(N_2)\\ &= n^4(q_1q_2)^2\left\{\left(1+\frac{1-q_1}{nq_1}\right)\left(1+\frac{1-q_2}{nq_2}\right)-1\right\} \text{ and so on.}\end{aligned}$$

Approximation of functions of moments

Let X_i, $i = 1, 2$ be two independent random variables with means μ_i and variances σ_i^2. Then it follows from a Taylor expansion that the function $g(X_1, X_2)$ has the approximate moments (Kotz and Jonsson (1985), p. 646)

$$\mathsf{E}\left[g(X_1, X_2)\right] \approx g(\mu_1, \mu_2) + \frac{1}{2}\left(\left[\frac{\partial^2 g}{\partial x_1^2}|\mu\right] \cdot \sigma_1^2 + \left[\frac{\partial^2 g}{\partial x_2^2}|\mu\right] \cdot \sigma_2^2 + 2\left[\frac{\partial^2 g}{\partial x_1 \partial x_2}|\mu\right] \cdot \sigma_{12}\right)$$

$$\mathsf{Var}\left[g(X_1, X_2)\right] \approx \left[\frac{\partial g}{\partial x_1}|\mu\right]^2 \cdot \sigma_1^2 + \left[\frac{\partial g}{\partial x_2}|\mu\right]^2 \cdot \sigma_2^2 + 2\left[\frac{\partial g}{\partial x_1}|\mu\right]\left[\frac{\partial g}{\partial x_2}|\mu\right] \cdot \sigma_{12} \tag{A4}$$

where all derivatives are evaluated at $\mu = (\mu_1, \mu_2)$. Also,

$$\mathsf{E}\left[g(X_i)\right] \approx g(\mu_i) + \frac{1}{2}\left[\frac{\partial^2 g}{\partial x_i^2}|\mu_i\right] \cdot \sigma_i^2 \text{ and } \mathsf{Var}\left[g(X_i)\right] \approx \left[\frac{\partial g}{\partial x_i}|\mu_i\right]^2 \cdot \sigma_i^2.$$

Legends to Figures

Figure 1: Calculation of the differences between the probability π with $Z_1 = 1$ and $Z_2 = 1$ in the independent and dependent case.

Figure 2: $\mathsf{Var}[\hat{\pi}]$ from (4) in the case with two dependent predictors (Z_1, Z_2), given that $n = 400$ and $p^{(1)} = \frac{1}{2}$.

Figure 3: $\mathsf{Var}[\hat{\pi}]$ from (4) in the case with two dependent predictors (Z_1, Z_2), given that $n = 400$ and $p^{(1)} = .10$.

Figure 4: $\mathsf{Var}[\hat{\pi}]$ from (4) in the case with two dependent predictors (Z_1, Z_2), given that $n = 400$ and $p^{(1)} = .90$.

Figure 5: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = \frac{1}{2}$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .05$.

Figure 6: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = .05$ and $q_{34}^{(0)}(\cdot) = .10$.

Figure 7: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = .05$ and $q_{34}^{(0)}(\cdot) = .20$.

Figure 8: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = .05$ and $q_{34}^{(0)}(\cdot) = .30$.

Figure 9: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .05$.

Figure 10: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .10$.

Figure 11: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .20$.

Figure 12: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .30$.

Figure 13: Predictive values for healthy (solid line) and non-healthy (dotted line) for various sample sizes.

Figure 14: $\mathsf{Var}[\hat{\pi}]$ as a function of number of independent sets of predictors.

Figure 1

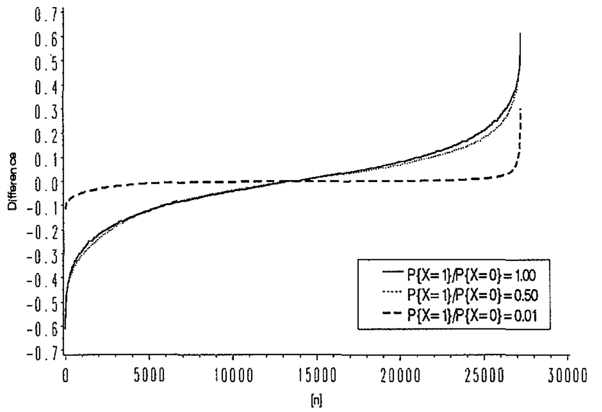


Figure 2

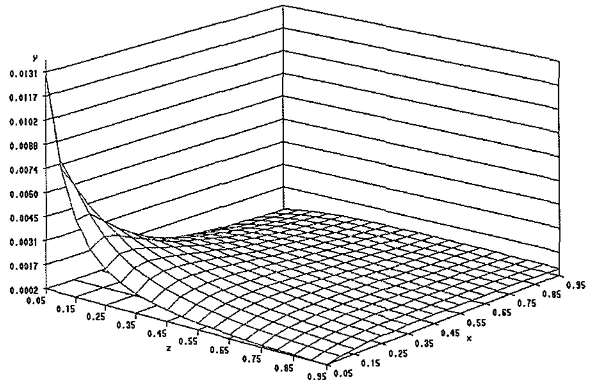


Figure 3

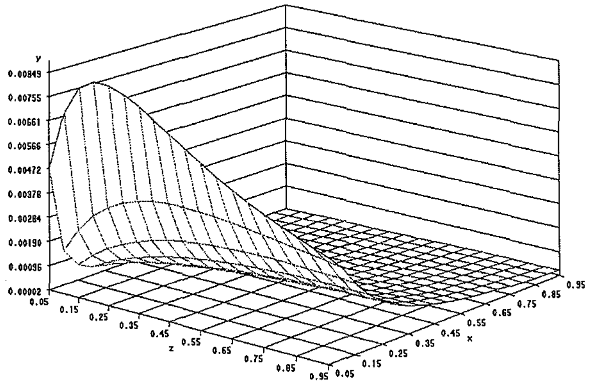


Figure 4

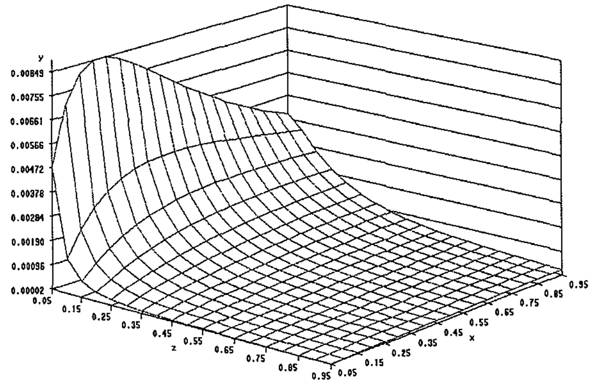


Figure 5

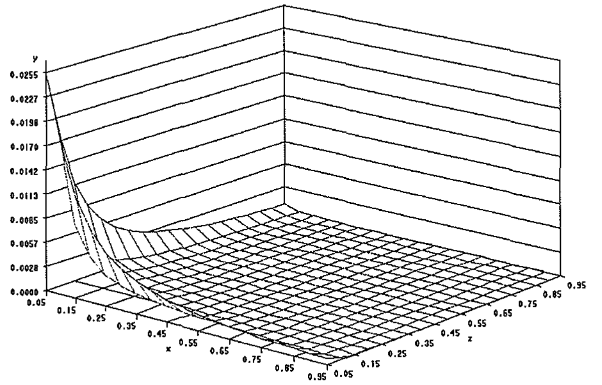


Figure 6

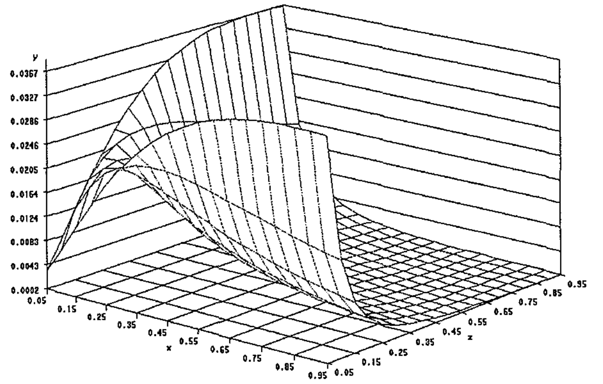


Figure 7

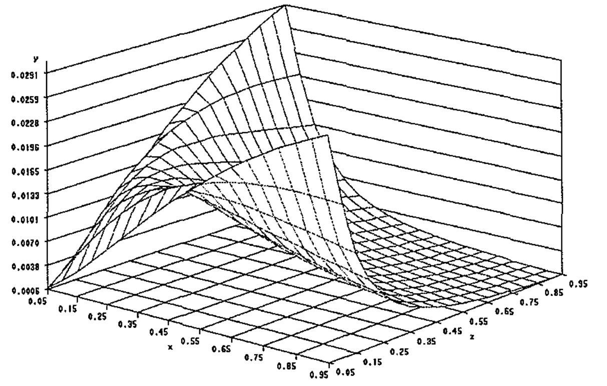


Figure 8

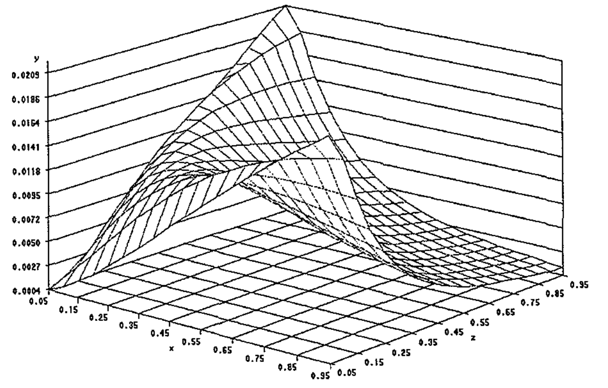


Figure 9

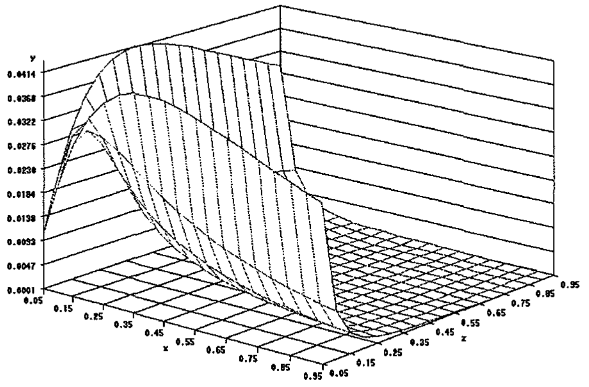


Figure 10

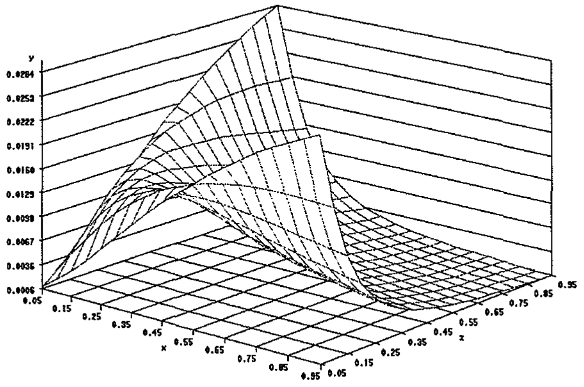


Figure 11

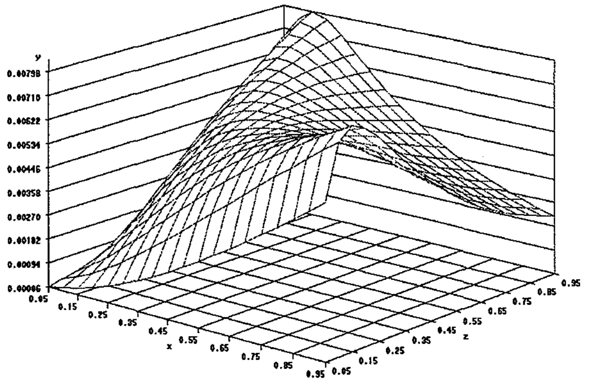


Figure 12

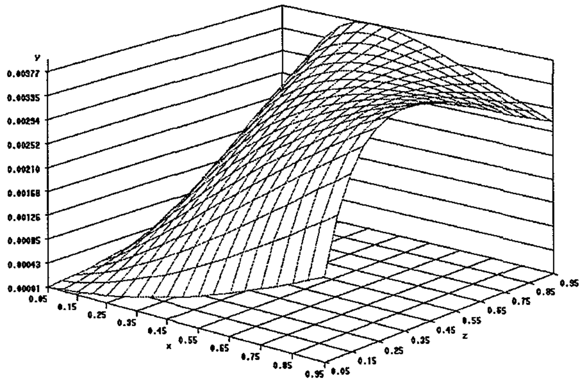


Figure 13

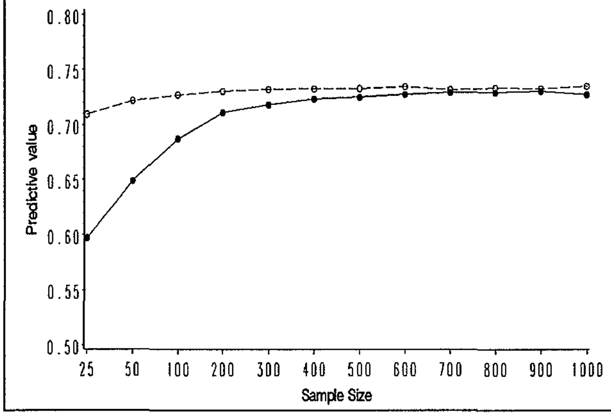
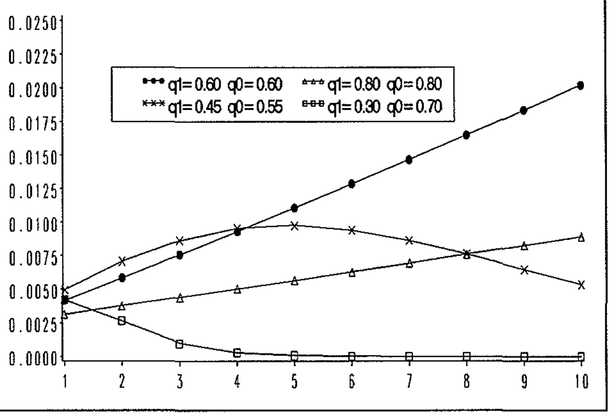


Figure 14



		Expected Length (z_1, z_2)				**Coverage Probability (%)** (z_1, z_2)			
CI	n	(1,1)	(1,0)	(0,1)	(0,0)	(1,1)	(1,0)	(0,1)	(0,0)
(11: *i*)	50	.29	.36	.36	.50	95	15	15	28
	100	.20	.60	.60	.73	95	40	40	59
	200	.14	.79	.78	.76	95	72	72	84
	400	.10	.70	.70	.58	95	89	89	92
	800	.07	.50	.50	.41	95	93	93	93
(11: *ii*)	50	.27	.84	.84	.82	95	64	64	78
	100	.20	.79	.79	.75	95	87	87	94
	200	.14	.71	.71	.63	95	97	97	97
	400	.10	.58	.58	.50	95	96	96	96
	800	.07	.45	.45	.38	95	96	96	96
(11: *iii*)	50	.29	1.69	1.69	1.55	95	63	63	77
	100	.20	1.44	1.43	1.23	95	85	85	92
	200	.14	1.08	1.08	.86	95	94	94	95
	400	.10	.73	.73	.59	95	96	95	96
	800	.07	.50	.50	.41	95	95	95	95
(11: *iv*)	50	.30	.94	.94	.92	97	39	39	53
	100	.21	.89	.89	.85	96	64	64	78
	200	.15	.81	.81	.73	96	87	86	94
	400	.10	.67	.67	.57	96	97	97	97
	800	.07	.50	.50	.42	95	97	97	97
(11: *v*)	50	.28	.84	1.06	.82	95	15	15	28
	100	.20	.80	.98	.76	95	40	40	60
	200	.14	.72	.84	.65	95	75	75	90
	400	.10	.59	.66	.51	95	95	95	97
	800	.07	.45	.48	.38	95	96	96	96

Table 2: Expected lengths and actual coverage probabilities (%) of the various CI's in (11): (*i*)-(*v*) for π, based on two dependent binary predictors. The q probabilities were $q^{(x)}(1,1) = .93$, $q^{(x)}(1,0) = .02$, $q^{(x)}(0,1) = .02$ and $q^{(x)}(0,0) = .03$, $x = 0,1$. The stipulated CI-level was 95%, and each figure was computed from 100,000 simulations.

	Expected Length: z_1,z_2,z_3,z_4															
Sample size, *n*	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	.42	.41	.68	.65	.48	.45	.69	.66	.66	.64	.76	.75	.61	.59	.74	.72
100	.29	.30	.64	.57	.37	.34	.65	.58	.59	.58	.72	.69	.54	.51	.70	.66
200	.20	.21	.58	.46	.28	.25	.57	.47	.51	.48	.65	.58	.44	.41	.62	.54
400	.13	.15	.48	.35	.21	.18	.46	.35	.39	.36	.51	.43	.30	.27	.47	.37
800	.10	.11	.37	.25	.15	.13	.33	.26	.27	.23	.33	.27	.17	.13	.26	.19
1600	.07	.08	.28	.18	.11	.09	.23	.18	.18	.14	.18	.17	.10	.07	.10	.09

	Coverage Probability (%): z_1,z_2,z_3,z_4															
Sample size, *n*	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	94	96	25	59	99	97	25	59	31	29	07	18	28	28	06	17
100	95	96	55	89	97	97	53	88	56	54	28	48	53	53	28	47
200	95	96	85	97	96	96	81	96	80	79	65	78	79	78	64	77
400	95	95	98	97	95	95	93	96	92	91	88	91	91	91	87	91
800	95	95	98	96	95	95	95	96	95	94	93	94	94	94	92	93
1600	95	95	96	95	95	95	95	95	95	95	94	95	95	95	94	94

Table 3: Expected length and actual coverage probabilities (%) of the various CI's in (12): (*i*) for π, based on two independent groups of dependent binary predictors (Z_1,Z_2) and (Z_3,Z_4). The *q* probabilities were $q_{12}^{(1)}(1,1)=.24$, $q_{12}^{(1)}(1,0)=.38$, $q_{12}^{(1)}(0,1)=.11$, $q_{12}^{(1)}(0,0)=.27$, $q_{12}^{(0)}(1,1)=.71$, $q_{12}^{(0)}(1,0)=.25$, $q_{12}^{(0)}(0,1)=.02$, $q_{12}^{(0)}(0,0)=.02$, $q_{34}^{(1)}(1,1)=.34$, $q_{34}^{(1)}(1,0)=.55$, $q_{34}^{(1)}(0,1)=.04$, $q_{34}^{(1)}(0,0)=.07$, $q_{34}^{(0)}(1,1)=.45$, $q_{12}^{(0)}(1,0)=.48$, $q_{34}^{(0)}(0,1)=.02$, $q_{34}^{(0)}(0,0)=.05$ and $p^{(1)}=.50$. The stipulated CI-level was 95%, and each figure was computed from 100,000 simulations.

	Expected Length: z_1,z_2,z_3,z_4															
Sample size, *n*	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	.38	.41	.80	.74	.54	.48	.83	.78	.79	.74	.92	.89	.64	.55	.82	.78
100	.27	.30	.75	.63	.40	.36	.75	.65	.66	.59	.80	.74	.46	.36	.63	.54
200	.19	.21	.66	.49	.29	.26	.61	.49	.49	.41	.57	.50	.29	.22	.37	.30
400	.13	.15	.53	.36	.21	.19	.45	.36	.34	.27	.33	.31	.19	.13	.18	.16
800	.09	.11	.40	.26	.15	.13	.32	.26	.24	.19	.20	.21	.12	.09	.10	.10
1600	.07	.08	.29	.18	.11	.09	.23	.19	.17	.13	.13	.14	.09	.06	.6	.07

Coverage Probability (%): z_1,z_2,z_3,z_4																
Sample size, *n*	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	96	96	25	59	96	96	25	59	36	36	9	22	36	36	10	23
100	96	96	55	89	95	96	55	89	60	60	34	55	60	60	34	56
200	95	95	84	97	95	95	84	97	84	84	72	84	84	84	73	84
400	95	95	96	96	95	95	96	96	95	95	94	96	95	95	94	96
800	95	95	96	95	95	95	96	95	96	96	96	96	96	96	96	96
1600	95	95	95	95	95	95	95	95	95	95	95	95	96	95	95	95

Table 4: Expected length and actual coverage probabilities (%) of the various CI's in (12): (*ii*) for π, based on two independent groups of dependent binary predictors (Z_1,Z_2) and (Z_3,Z_4). The same *q*-probabilities as in Table 3 were used. The stipulated CI-level was 95%, and each figure was computed from 100,000 simulations.

Prediction of Work Resumption Among Men and Women with Lower Back- and Neck Pain in a Swedish Population

by Anders Persson

Department of Statistics, Göteborg University, Sweden

Abstract

An approach based on Bayes theorem is used to predict the binary outcome of work resumption *X*, where $X = 1$ if no work resumption and $X = 0$ otherwise, given a vector of discrete predictors **Z** for men and women with lower back- and neck pain in a Swedish population. In this application the predictors have a complex dependency structure. Hierarchical cluster analysis is used to create independent groups of dependent predictors such that predictors within groups are dependent while predictors in different groups are independent. The main purpose is to estimate the probability $P(X = 1|\mathbf{z})$ and to calculate confidence intervals for this probability. Based on these estimates one may decide whether a given person should be predicted as healthy or as non-healthy, and predictive values are calculated in order to evaluate of the performance of the prediction analysis. The results are compared with the frequently used ordinary logistic regression method without interactions. It is found that ignoring the correlations between the predictors may give seriously misleading results. Also, the problem with missing values is discussed.

Key words: Confidence intervals; Hierarchical cluster analysis; Logistic regression; Prediction; Predictive value; Work resumption.

1 INTRODUCTION

In many applications the aim is to predict a binary outcome given the value of a set of predictor variables. A commonly used method for this situation is ordinary logistic regression (Cox (1970); Hosmer and Lemeshow (1989); McCullagh and Nelder (1989) and Neter et al. (1996)). In many applications, the predictors have a complex dependency structure, which might be difficult to capture with the logistic model. Although the use of interaction terms works well in a logistic model with few predictors, problems may arise when there are many predictors. The reason for this is that there is a total of $2^k - 1$ β-parameters to estimate if all interactions are included. For obvious reasons it is almost impossible to include all interactions if there are many predictors.

In this paper we apply a method suggested by Jonsson and Persson (2002) which is based on Bayes theorem to predict the outcome variable 'work resumption' $(X = 0)$ and 'no work resumption' $(X = 1)$ among men and women with back- and neck pain diagnosis, conditional on the values of a discrete vector of predictors **Z**.

This paper is motivated by the fact that the number of long-term sick-listed individuals has been increasing persistently in Sweden and in many other countries. Back- and neck pain is one of the most frequently cases behind long-term sick-listing (Bergendorff et al. (1997) and Hansson and Hansson (1999)). Since the middle of the 80s the National Social Insurance Board (RFV) has conducted studies to identify important factors affecting health state improvement and work resumption. Due to increased efforts on economic and personal resources, including interventions to improve the propensity of work resumptions, it has resulted in amount of positive changes since the beginning of the 90s (Riksförsäkringsverket (1995) and Persson and Tasiran (2001)). But, during the period 1997-2001 the numbers of individuals who have been sick-listed longer than 365 days have increased from 75,000 to 120,000. The relative increase during the 4-year period have been about 30% per year and the number of earlier retirements/temporary disability pensions have increased from 423,000 to 450,000. Including waiting period, sick pay, sickness allowance and earlier retirements/temporary disability pensions it corresponds to 800,000 full-time annual jobs or 14 percent of the population at the ages

18-65. The associated costs for health insurance are 108 billion SEK according to the state budget 2002 (SOU (2002)).

A sample consisting of 1575 full-time working employed was available for the analyses. Four sub-groups were of special interest: men and women with back- or neck pain diagnosis, and were treated separately. The individuals in the sample were followed-up during a 2-year period and predictions were possible at 90 days, 1 year and 2 years after sick-listing, respectively.

The process of prediction proceeded in the following two stages. In a first step, the probabilities $\pi = P(X = 1|\mathbf{z})$ were estimated and confidence intervals were calculated for each probability. In a second step, new subjects were sampled sequentially from the same population by simulations to make predictions of 'no work resumption' given the values of a set of predictors based on the estimates in the first step.

The plan of the paper is as follows. Section 2 starts with a brief description of the material. In Section 3 the statistical methods are described. Section 4 deals with estimation of $P(X = 1|\mathbf{z})$ and calculation of confidence intervals for these probabilities. This section ends with prediction of 'no work resumption', and presents measures for prediction ability such as predictive values. Finally, in Section 5, some concluding remarks are given.

2 Material

A sample of 1575 full-time working employed sick-listed for at least 28 days because of lower back- or neck pain diagnosis followed-up during a two-year period was available for the analyses. Data were collected by the National Social Insurance Board (RFV) sequentially during the period November 1994 until October 1995 represented by 5 different counties of Sweden; Stockholm, Kristianstad, Västmanland, Västernorrland and Göteborg. Three time points were of special interest: 90 days, 1 year and 2 years after sick-listing. Individuals with both lower back- and neck pain diagnosis (240) were excluded from the analyses due to difficulties with confounding effects. Each of these 4 sub-groups (by sex and diagnosis) was treated separately due to large differences between their patterns of work resumption. For a detailed description of the material see Bergendorff et al. (1997); Bergendorff et al. (2001); Riksförsäkringsverket och Sahlgrenska universitetssjukhuset (1997) and Hansson and Hansson (1999).

Unfortunately, the data quality was rather low since there were considerable amounts of missing values on some predictor variables (see discussion in Section 4.1). Furthermore, only 5 counties participated in the study. Hence, the results were not representative for the whole population of Sweden.

Sometimes the term 'healthy' and 'non-healthy' will be used for simplicity rather than 'work resumption' and 'no work resumption', respectively. The state 'healthy' was defined as a sick-listed person who has become able to work. A person, who was fully or partially sick-listed, early retirement or entitled to temporary disability pension, was defined as a 'non-healthy' person (Bergendorff et al. (2001)). Occasionally, we use the abbreviation MB90, MB1Y, MN90, MN1Y, WB90, WB1Y, WN90 and WN1Y, where M=men, W=women, B=back pain, N=neck pain. 90=90 days and 1Y=1 year.

Baseline characteristics. *Sex*, *Age* (Z_1), *Diagnosis* and *County*. There were a total of 883 females (56%) and 692 males at the ages 18-59. The mean(SD) age was 42(10) years for all groups. In the analyses, Age was dichotomized where Age=1 if a person was older than 31 years and 0 if a person was younger than 31 years. High age was a positive factor for 'no work resumption' in all groups except for women with lower back pain diagnosis.

Table 1 below shows the prevalence in the sub-groups at 90 days, 1 year and 2 year after sick-listing. People with lower back pain recovered faster than those with neck pain. Among persons with lower back pain there were 42 percent healthy within 90 days, 79 percent within 1 year and 87 percent within 2 years. The corresponding figures for people with neck pain were 39, 73 and 81 percent, respectively. Men with lower back pain recovered faster than women with the same diagnosis, while there was no significant difference between men and women with neck pain (Bergendorff et al. (2001)).

	90 days	1 year	2 years
Men/Back	0.54	0.17	0.11
Men/Neck	0.60	0.30	0.21
Women/Back	0.63	0.24	0.15
Women/Neck	0.63	0.25	0.18

Table 1: Prevalence's at 90 days, 1 year and 2 years after sick-listing.

There was a strong connection between sex and diagnosis. Table 2 shows that men suffered more frequently from back problems (79%) as compared to women (63%), while women suffered more frequently from neck problems (37%) as compared to the men (21%). The diagnoses varied between the counties in the material. Table 3 below shows the distribution of lower back- and neck diagnosis in the 5 counties. Lower back pain was the most frequent cause of sick-listing in Stockholm (73%) while neck pain was most frequent in Västmanland (36%).

County	**Men/Back**	**Men/Neck**	**Women/Back**	**Women/Neck**
Stockholm	169	41	154	76
Kristianstad	87	30	88	56
Västmanland	64	23	74	56
Västernorrland	76	23	104	52
Göteborg	149	30	132	91
Total	**545**	**147**	**552**	**331**

Table 2: Number of cases of sick-listing by county, sex and diagnosis.

County	**Back (%)**	**Neck (%)**
Stockholm	**73**	27
Kristianstad	67	33
Västmanland	64	**36**
Västernorrland	71	29
Göteborg	70	30

Table 3: Distribution of lower back- and neck pain diagnosis in the 5 counties.

Socioeconomic factors. *Education* (Z_2), *Ethnicity* (Z_3) and *Household income* (Z_4). Education was defined on a 3 level ordinal scale with 1 as lowest and 3 as highest degree of education. Level 1 and 2 representing low education (=1) and level 3=high education (=0). Ethnicity is a 20 level nominal variable where 1=Swedish and 2-20 representing non-Swedish (=0). Finally, Household income was a continuous variable ranging from 900 to 175,000 SEK, dichotomized as 1 if > 7000 SEK and 0 otherwise.

Psychical working environment. *Demand* (Z_5), *Control* (Z_6), *Strain* (Z_7) and *Attitude* (Z_8). Demand was expressed as self experienced demands on their place of work, scaled 25 (low)-100 (high), where 25-70 was defined as low (=0) and 70-100 as high (=1). Control is the possibility of affecting their own working environment scaled 25 (low)-100 (high), where 25-70 was defined as low (=1) and 70-100 as high (=0). Strain is simply the ratio between Demand and Control, where 0.25-0.84 was defined as low (=0) and $\geq$ 0.84 as high (=1). Attitude was measured on a scale 3 (low)-9 (high) where 0-4 was defined as low (=0) and $\geq$ 5 as high (=1).

Physical working environment. *Inconvenient working environment* (Z_9), *Heavy lifts* (Z_{10}) and *Suitable working tasks* (Z_{11}). By 'Inconvenient working environment' and 'Heavy lifts' we mean 4 level variable ranging from 1 (yes, often) to 4 (no, never), where 1-2 was defined as yes (=1) and 3-4 as no (=0). Finally, by 'Suitable working tasks' is meant that the employer was willing to adjust the working tasks in agreement with the individual's state of health, where 1=no and 0=yes.

Family and social networks. *Sick-listing in the family* (Z_{12}), *Temporary disability pension/early retirement in the family* (Z_{13}) and *Offered temporary disability pension/early retirement* (Z_{14}). All variables were dichotomous where 1=yes and 0=no.

Health state. *Work ability* (Z_{15}), *Comorbidity* (Z_{16}) and *Smoking* (Z_{17}). Working ability was subjectively assessed on a scale ranking from 1 (low) to 10 (high), where 1-4 was defined as bad working ability and 5-10 as good working ability. By Comorbidity we mean that the individual has other diseases than lower back- or neck pain, where 1-2 was defined as no (=0) and 3 as yes (=1). Smoking was a 3 level variable defined as yes or never smoked (=1) and quit smoking (=0).

Administrative interventions. The presence of *Complete rehabilitation plan* (Z_{18}) was a dichotomous variable defined as 1=yes and 0=no.

Predictor	**MB (%)** $n = 545$	**MN (%)** $n = 147$	**WB (%)** $n = 552$	**WN (%)** $n = 331$	**p-value**
Age	84	82	86	83	.35
Education	90	94	93	95	.11
Ethnicity	19	21	14	26	<.01
Household income	96	99	91	91	<.01
Demand	48	63	56	66	<.01
Control	34	32	48	56	<.01
Strain	62	74	73	83	<.01
Attitude	84	83	90	87	.07
Inconvenient working environment	85	93	85	89	.17
Heavy lifts	82	82	80	76	.48
Suitable working tasks	50	64	62	63	.02
Sick-listing in the family	13	13	9	8	.15
TDP/ER in the family	17	15	14	12	.31
Offered TDP/ER	11	24	12	17	.01
Work ability	51	50	53	45	.26
Comorbidity	8	15	8	8	.13
Smoking	67	64	74	75	.05
Rehabilitation plan	18	21	26	24	.02

Table 4: Descriptive statistics and χ^2-test of equal proportions between the 4 sub-groups. The proportions in the table are given that all predictors equals to 1 (see definitions in Table A1 in Appendix). The abbreviation TDP/ER denotes Temporary Disability Pension/Early Retirement.

3 Statistical Methods

A method based on Bayes theorem for predicting a binary outcome $X = 0,1$ given the values of a vector of discrete predictors **Z**, suggested by Jonsson and Persson (2002) is used for the analyses. The probability π was estimated according to (3) and 95% confidence limits according to (12:(*i*)) in the latter work. At baseline i.e. after 28 days of sick-listing a large set of predictors was available from the material. In a previous study (Bergendorff et al. (2001)) a list of potential predictors has been proposed for prediction of work resumption among men and women with lower back- and neck pain (see Table

5). The predictors were chosen on basis of probability plots. In a second step, a hierarchical clustering method (Anderberg (1973) and Jobson (1992)) have been used to create independent groups of dependent predictors both given $X = 1$ and $X = 0$. That is, for a given value of X the purpose is to identify groups of predictors such that predictors within groups are dependent but at the same time are independent of predictors in other groups. Consequently, it is not necessarily the same predictors in the groups given $X = 1$ and $X = 0$, respectively. In addition to the cluster analysis Pearson's correlation coefficient have been calculated between the predictors both given $X = 1$ and $X = 0$ to examine the dependency structure in detail. Although a χ^2-test of independence in a 2×2 contingency table may be sufficient, the correlation coefficient is perhaps a better descriptive measure of association between the predictors. In fact, the χ^2-test and Pearson's correlation coefficient are related by $r = \{n^{-1}X^2\}^{1/2}$, where r is the correlation coefficient, X^2 is the value of the chi-square statistic and n is the number of observations.

	Men				**Women**			
	Back		**Neck**		**Back**		**Neck**	
Predictor	**90d**	**1y**	**90d**	**1y**	**90d**	**1y**	**90d**	**1y**
Z_1	X	X	X	X				
Z_2		X	(X)	X				X
Z_3	X	(X)						
Z_4							X	X
Z_5	X	(X)	(X)	(X)	X	X	X	(X)
Z_6			X					
Z_7			(X)				X	
Z_8			(X)				(X)	
Z_9	(X)	(X)					(X)	(X)
Z_{10}	(X)					(X)		(X)
Z_{11}	X	X		(X)			(X)	(X)
Z_{12}	X							
Z_{13}			(X)	X				
Z_{14}			X	(X)		(X)	(X)	(X)
Z_{15}	X	X	X	X	X	X	X	X
Z_{16}	X	X	X	X	X	X	X	X
Z_{17}			(X)	X				
Z_{18}	X	(X)	(X)	(X)	X	X	X	X

Table 5: Potential predictors for prediction analysis at 90 days and 1 year after sick-listing (Bergendorff et al. (2001)). Predictors marked with (X) were not included in the models. See Table A1 in Appendix for labels to the predictors.

One possibility to test whether a predictor has a significant effect on the outcome is a stepwise logistic regression. However, this method cannot be used for testing whether the predictors are dependent or not conditionally on $X=1$ and $X=0$. This follows easily from the illustrations in (Jonsson and Persson (2002), p. 7). Furthermore, with 8 predictors there are up to 255 β–coefficients to be tested in a pre-test, and this give rise to inferential problems. But, there is another possibility that we might consider. Let $\mathbf{z}_1=(z_i=1\,,\mathbf{z}_r)$ be the vector of all predictors with the constraint that the *i*th predictor takes on the value 1 and $\mathbf{z}_0=(z_i=0\,,\mathbf{z}_r)$ that the *i*th predictor takes on the value 0, where $\mathbf{z}_r$ is a subset of $\mathbf{z}$ when the *i*th predictor is excluded. The effect of the predictor Z_i given $\mathbf{z}_r$ can be expressed as the estimated differences $\hat{\delta}=\hat{\pi}_1-\hat{\pi}_0$, where $\hat{\pi}_1=\hat{P}(X=1|\,\mathbf{z}_1)$ and $\hat{\pi}_0=\hat{P}(X=1|\,\mathbf{z}_0)$. For example, if $\mathbf{z}=(z_1,z_2)$ and $\mathbf{z}_r=(z_2)$ then $\mathbf{z}_1=(z_1=1\,,z_2)$ and $\mathbf{z}_0=(z_1=0\,,z_2)$. Since z_2 can take on the values 0 or 1 there are 2 possible outcomes for $\hat{\delta}$, $\hat{\pi}_1$ and $\hat{\pi}_0$, respectively. The difference δ is estimable if and only if there are observations on both z_1 and z_2. Let n' be the number of estimable $\delta's$. Then, $\max\{n'\}=2^{k-1}$, where *k* is the number of predictors. We want to test the hypothesis $H_0:\delta=0$ given that the predictors $\mathbf{z}_r$ are in the model, against the alternative $H_A:\delta\neq 0$. It can be performed in many ways. With few estimable $\delta's$ a Sign test may be appropriate. If the number of observations is sufficiently large and normal distribution of the $\hat{\delta}'s$'s can be assumed, a test based on normality may be better, or if the distribution is at least symmetric a Wilcoxon Signed Rank test may be appropriate (Altman (1991)). These tests require a large set of predictors and very few missing values. For example, if there are two predictors in the model, there are only two differences to calculate. This will be further explained in Section 4.1, and examples will be given in Table 9.

4 PREDICTION

This section is devoted to prediction of the binary outcome $X=$'work resumption' conditional on the values of a vector of discrete predictors **Z**. We are primarily interested in predicting 'no work resumption' $(X=1)$. The reason for this is that among non-healthy persons it was desirable to find characteristics such that appropriate interventions e.g.

rehabilitation actions that gain work resumption can be taken as soon as possible after sick-listing. Predictions were made 90 days and 1 year after sick-listing, respectively. A detailed discussion is given in Section 4.1 for men with back pain (90 days) only. But, in Section 4.2 we summarize and compare the prediction results from the remaining sub-groups as well.

4.1 Men with Lower Back Pain (90 days)

There were 545 men with lower back pain diagnosis available for the analysis. Initially, there were 10 potential predictors of interest (see Table 5), but these have been reduced to 8 predictors. There were considerable amounts of missing values for most of the predictors (see Table A1 in Appendix). For example, the predictor 'Suitable working tasks' had 333 (61%) missing values. With k binary predictors there are 2^k possible outcomes for π. For example, with 8 predictors there are 256 various outcomes that require a rather large sample size and few missing values. The sample size needed for estimation of the $\pi's$ depend on the distribution of the cell frequencies (Jonsson and Persson (2002)).

Table 6 shows the dependency structure among the 8 chosen predictors given 'no work resumption' $(X=1)$ and 'work resumption' $(X=0)$, respectively. Note that there were not the same predictors in the groups given $X=1$ and given $X=0$. That is, the composition of predictors across groups affecting the probability of 'no work resumption' is different from the probability of 'work resumption'. The following dependency structure was obtained from the hierarchical cluster analysis. Note that it is not the same predictors in Table 6 as in the simulation example in Section 5.1 in Jonsson and Persson (2002).

Group	Predictors associated with 'no work resumption' $(X = 1)$
1	Rehabilitation plan (Z_{18}), Demand (Z_5), Suitable working tasks (Z_{11})
2	Sick-listing in the family (Z_{12}), Ethnicity (Z_3)
3	Comorbidity (Z_{16}), Work ability (Z_{15}) Age (Z_1)

Group	Predictors associated with 'work resumption' $(X = 0)$
1	Comorbidity (Z_{16}), Demand (Z_5), Suitable working tasks (Z_{11}), Ethnicity (Z_3)
2	Rehabilitation plan (Z_{18}), Work ability (Z_{15}) Age (Z_1), Sick-listing in the family (Z_{12})

Table 6: Result of hierarchical cluster analysis for men with lower back pain (90 days).

Table 7 and 8 shows the correlations between pairs of predictors given $X = 1$ and $X = 0$. It is seen that the hierarchical clustering method to some extent agrees with the correlation coefficients between pairs. But, from Table 7 it is seen that Age (Z_1) in group 3 given $X = 1$ is pairwise independent of Work ability (Z_{15}) and Comorbidity (Z_{16}) with correlations .00 and -.05, respectively. However, Age (Z_1) is at the same time independent of every predictor in the group 1 and 2. Furthermore, in group 2 given $X = 0$, Table 8 shows that Rehabilitation plan (Z_{18}) is pairwise independent of Age (Z_1), Sick-listing in the family (Z_{12}) and Work ability (Z_{15}) with correlations .01, .08 and .06, respectively. But, Rehabilitation plan (Z_{18}) is at the same time independent of every predictor in group 1. It should be noticed that pairwise independency is not the same as simultaneously independency.

	Z_1	Z_3	Z_5	Z_9	Z_{10}	Z_{11}	Z_{12}	Z_{15}	Z_{16}	Z_{18}
Z_1	1	-	-	-	-	-	-	-	-	-
Z_3	.07	1	-	-	-	-	-	-	-	-
Z_5	-.04	.11	1	-	-	-	-	-	-	-
Z_9	-.05	.13	.09	1	-	-	-	-	-	-
Z_{10}	.06	.05	.10	**.48**	1	-	-	-	-	-
Z_{11}	.07	-.01	**.36**	**.27**	**.33**	1	-	-	-	-
Z_{12}	.03	**-.15**	-.04	**-.16**	.01	-.12	1	-	-	-
Z_{15}	-.05	.10	.14	.08	-.05	**.25**	.01	1	-	-
Z_{16}	.00	**.23**	.05	.02	-.05	.09	-.03	**.15**	1	-
Z_{18}	.01	-.03	**.19**	**.16**	.14	**.25**	.01	**.14**	.12	1

Table 7: Correlation matrix for predictors among men with lower back pain diagnosis (90 days) associated with 'no work resumption' $(X = 1)$. Significant correlations (5%) are marked with bold type $(n^{(1)} = 295)$.

	Z_1	Z_3	Z_5	Z_9	Z_{10}	Z_{11}	Z_{12}	Z_{15}	Z_{16}	Z_{18}
Z_1	1	-	-	-	-	-	-	-	-	-
Z_3	.15	1	-	-	-	-	-	-	-	-
Z_5	.06	.09	1	-	-	-	-	-	-	-
Z_9	-.04	-.06	.07	1	-	-	-	-	-	-
Z_{10}	-.15	-.17	.05	**.44**	1	-	-	-	-	-
Z_{11}	-.02	.12	**.23**	**.26**	.17	1	-	-	-	-
Z_{12}	**-.20**	.05	.07	.11	.04	**.23**	1	-	-	-
Z_{15}	-.09	.10	.01	.06	-.02	.07	**.17**	1	-	-
Z_{16}	.08	**.29**	.14	-.01	.02	**.19**	.07	**.20**	1	-
Z_{18}	.01	-.08	.06	.10	.12	.01	.08	**.23**	-.04	1

Table 8: Correlation matrix for predictors among men with lower back pain diagnosis (90 days) associated with 'work resumption' $(X = 0)$. Significant correlations (5%) are marked with bold type $\left(n^{(0)} = 250\right)$.

Sparse contingency tables often contain cells having zero frequency counts or missing values. Cells for which a nonzero count is impossible because of the design of the study are sometimes referred to as structural zeros. In this application, however, we are only concerned with missing values and sampling zeros i.e. nonzero counts are possible, but a zero occurs because of random variation. Sampling zeros are especially likely to arise when the sample is small and the contingency table has many cells (Agresti (1991)). Out of the 545 observations there were 186 observations available for prediction and only 50 (20%) of the 256 probabilities were estimable due to missing values and sampling zeros. It means that if new individuals are sampled from the same population in the same way as in the original survey, it is likely that some individuals have values of the predictors such that predictions for those subjects are not possible. The numbers of missing values for each predictor are presented in Table A1 in Appendix.

The separate effect for each predictor in the model is illustrated in Table 9 with the δ-test (see also Figures 9-16 for plots of the $\hat{\delta}'s$ for each predictor). The results in Table 9 show that individuals with complete rehabilitation plan, bad work ability, sick-listing in the family and people who did not have suitable working tasks had higher probability of 'no work resumption'. But, Age, Comorbidity, Demand and Ethnicity did not show any significant differences indicating that these should be excluded from the model. However, due to the fact that there are very few estimable $\delta's$ (n') the reliability of the test result may be questionable. The reason for that n' is relatively small compared to the 50 estimable probabilities is that the test require values of every predictor in the vector $\mathbf{z}_r$ both given $Z_i = 1$ and $Z_i = 0$. Otherwise, δ is not estimable for that combination of **Z**. With 8 predictors the maximal value of n' is 128. None of the predictors in Table 9 has a value of n' greater than 20 and the value for Comorbidity is as low as 7.

Predictor	n'	**Mean($\hat{\delta}$)**	**Median($\hat{\delta}$)**	**Std. dev($\hat{\delta}$)**	**p-value**
Age (Z_1)	9	.037	.008	.102	.16
Rehab. plan (Z_{18})	11	.384	.291	.234	<.01
Comorbidity (Z_{16})	7	.018	.060	.202	.94
Work ability (Z_{15})	17	.175	.223	.199	<.01
Demand (Z_5)	17	-.041	-.064	.096	.09
Sick-listing in the family (Z_{12})	10	.243	.241	.200	<.01
Suitable working tasks (Z_{11})	18	.194	.178	.117	<.01
Ethnicity (Z_3)	10	.093	.114	.221	.19

Table 9: Descriptive statistics and a Sign test of the differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for testing if the predictors have an effect on the outcome variable.

Table 10 shows the frequencies of predicted work resumption versus the true state.

		True state		
		Healthy	Non-healthy	
Predicted	Healthy	81	33	114
State	Non-healthy	15	57	72
		96	90	186

Table 10: Predicted and true state of work resumption for men with lower back pain 90 days. Out of the 545 individuals only 186 observations were available for prediction due to missing values on the predictor variables.

In order to evaluate the prediction ability, simulations have been used to sample new individuals (100,000) from the same population. The prediction ability was evaluated by predictive values, relative predictive values and proportion of correct classifications. But the predictive value could in some cases be misleading without reference to the prevalence. For example, a predictive value of .92 and prevalence .90 is obviously not as good as if the prevalence was .20, say. Therefore, it seems more reasonable to use relative predictive values. The latter show the relative gain in predicting the outcome rather than simply guessing the outcome in accordance with the prevalence. From Table

1 it can be seen that the prevalence at 90 days was 0.54. The predictive value for 'no work resumption' and 'work resumption' was 0.76 and 0.74, respectively, with the corresponding relative predictive values 42% and 37%. The proportion of correct classifications was 0.75 (see also Table 12 for comparisons with the remaining sub-groups).

We recall that various values of $\mathbf{Z}$ give different values of π. Figure 1 shows the ordered values of $\hat{\pi}$ and associated confidence limits. Since, it is not possible from the figure to identify the values of the predictors represented by the index variable on the x-axis, Table A2 in Appendix presents all estimable π's, confidence limits and $\hat{V}[\hat{\pi}]$. The following examples illustrate how to interpret the results.

Example 1. Mr. A is a Swedish man older than 31 years entitled to rehabilitation plan. He has other diseases than lower back pain, bad working ability and he experience high demand at his place of work, where his working tasks are not appropriate for him. Furthermore, he has no cases of sick-listing in his family. Mr. A has a probability of 'no work resumption' equal to 0.97 with confidence limits (0.38; 0.99).

Example 2. Mr. B is a Swedish man older than 31 years with no rehabilitation plan. He has no other diseases than lower back pain, good working ability and he does not experience high demand at his place of work, where his working tasks are appropriate for him. Furthermore, he has no cases of sick-listing in his family. Mr. B has a probability of 'no work resumption' equal to 0.14 with confidence limits (0.09; 0.26).

4.2 Comparison of the Prediction Results for All Sub-groups

Table 11 shows the proportion of $\hat{\pi} \leq \frac{1}{2}$ and $\hat{\pi} > \frac{1}{2}$ in all the 4 sub-groups. It is seen that the proportion of $\hat{\pi} > \frac{1}{2}$ is rather high at 90 days for all groups and low after 1 year. For a detailed examination of the estimated probabilities, see Figures 1-8 and Table A2, which show the ordered values of $\hat{\pi}$ and associated confidence limits.

The results obtained from the prediction analysis for all sub-groups are showed in Table 11 and 12. It is seen from Table 11 that the proportion of estimable probabilities is rather low fore some groups and high for others. But, there is different number of predictors in the sub-groups. Of course, it easier to obtain a higher proportion of estimable probabilities with fewer predictors.

Group	n	No. of obs. for prediction	No. of predictors	No. of probabilities to estimate	No. of estimable probabilities	Proportion of $\hat{\pi} > \frac{1}{2}$
MB90	545	186	8	256	50	0.58
MB1Y	545	161	6	64	23	0.17
MN90	147	47	6	64	9	0.56
MN1Y	147	80	6	64	21	0.24
WB90	552	309	4	16	16	0.87
WB1Y	552	303	5	32	26	0.42
WN90	331	111	6	64	16	0.56
WN1Y	331	138	5	32	14	0.36

Table 11: Basic statistics for all sub-groups, separately.

In Table 12 below it can be seen that the prediction ability after 1 year is better performed as compared to 90 days. But, there are no differences in prediction ability between men and women and between lower back and neck pain diagnosis.

Group	Prevalence	Predictive value non-healthy	Relative predictive value non-healthy	Predictive value healthy	Relative predictive value healthy	Proportion of correct classified
MB90	0.54	0.76	42%	0.74	37%	0.75
MB1Y	0.17	0.59	244%	0.91	435%	0.86
MN90	0.60	0.68	14%	0.82	37%	0.70
MN1Y	0.30	0.68	127%	0.81	170%	0.78
WB90	0.63	0.81	28%	0.63	0%	0.73
WB1Y	0.24	0.65	169%	0.86	258%	0.82
WN90	0.63	0.81	29%	0.62	-2%	0.73
WN1Y	0.25	0.70	199%	0.85	240%	0.82

Table 12: Prediction ability for all sub-groups, separately.

5 DISCUSSION

This paper is an application of an approach suggested by Jonsson and Persson (2002) based on Bayes theorem. The aim is to predict the outcome 'no work resumption' conditionally on the values of a set of discrete predictors, and also to make CI statements. It is emphasized that problems may arise when some of the predictors have a considerable amount of missing values. The consequences of getting missing values may be serious. First, the number of observations available for prediction and the number of estimable probabilities decreases. Secondly, if new subjects are sampled sequentially from the same population, it is likely that we obtain individuals with values on the vector of predictors such that π is not estimable. The material in this application contains many missing values, which would have justified the use of fewer predictors in the model. In the latter case we would have obtained relatively more observations for prediction. Fewer predictors do not necessarily alter the prediction ability. Also, the number of parameters to estimate increases dramatically as the number of predictors in the model increases.

The proposed method works very well for most cases. Correct specification of the dependency structure is a matter of crucial importance. The assumption of independent predictors when they in fact are correlated may lead to seriously misleading results concerning bias and variance (Jonsson and Persson (2002)). For example, for men with lower back pain (90 days) where $Z_1 = Z_3 = Z_5 = Z_{12} = Z_{18} = 0$ and $Z_{11} = Z_{15} = Z_{16} = 1$ we obtain $\hat{\pi} = 0.76$ for the Bayes approach with a corresponding estimate of 0.29 for an ordinary logistic regression model. In this paper we have used the decision rule; if $\pi > \frac{1}{2}$ then a given subject is predicted 'no work resumption' and if $\pi \leq \frac{1}{2}$ then the subject is predicted 'work resumption'. The choice of the limit is somewhat arbitrary, but in a real life situation the estimated probability will be used in conjunction with other sources of information about the sick-listed person to reach a decision whether e.g. interventions should be taken.

A test for detecting separate variable effects in the model was suggested. Since the test depends on the number of predictors in the model it seems inappropriate to use such a test for materials with many missing value and few potential predictor variables.

The results of the predictions showed that the prediction ability after 1 year was better performed as compared to 90 days, as measured by relative predictive values. But, there were no differences in prediction ability between men and women and between lower back- and neck pain diagnosis.

ACKNOWLEDGMENTS

I would like to acknowledge a special appreciation to my supervisor Ph. D. Robert Jonsson for advice and encouraging support during the work of this paper. Thanks to Christian Sonesson for constructive comments and suggestions on earlier versions of the manuscript.

REFERENCES

[1] Agresti, A. (1990) *Categorical Data Analysis*. New York: John Wiley & Sons.

[2] Altman, D.G. (1991) *Practical Statistics for Medical Research*. London: Chapman and Hall.

[3] Anderberg, M.R. (1973) *Cluster Analysis for Applications*. New York: Academic Press.

[4] Bergendorff, S. Hansson, E., Hansson, T. and Jonsson, R. (2001) (In Swedish) *Vad kan förutsäga utfallet av en sjukskrivning?* Rygg och Nacke 8. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[5] Cox, D.R. (1970) *Analysis of Binary Data*. London: Chapman and Hall.

[6] Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression*. New York: Wiley.

[7] Jobson, J.D. (1992) *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods*. New York: Springer-Verlag.

[8] Jonsson, R. and Persson, A. (2002) *Bayes Prediction of Binary Outcomes Based on Correlated Discrete Predictors*. Research report 2002:3, Department of Statistics, Göteborg University.

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

[10] Neter et al. (1996) *Applied Linear Regression Models* (3rd ed.). London: Irwin.

[11] Persson, A. and Tasiran, A.C. (2001) *Analysis of Spatial Effects on Long-term Sicknesses*. Seminar paper 2001:5, Department of Statistics, Göteborg University.

[12] Riksförsäkringsverket (1995) (In Swedish) *RIKS-LS - en undersökning om långvarig sjukskrivning och rehabilitering*. RFV REDOVISAR 1995:20. Stockholm: Riksförsäkringsverket.

[13] Bergendorff, S. Hansson, E., Hansson, T., Palmer, E., Westin, M. and Zetterberg, C. (1997) (In Swedish) *Projektbeskrivning och undersökningsgrupp.* Rygg och Nacke 1. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[14] Riksförsäkringsverket och Sahlgrenska universitetssjukhuset (1997) (In Swedish) *Enkäter till undersökningsgruppen och försäkringskassan.* Rygg och Nacke 2. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[15] Hansson, E. and Hansson, T. (1999) (In Swedish) *Medicinsk åtgärder för sjukskrivna för rygg- och nackbesvär.* Rygg och Nacke 3. Stockholm: Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset.

[16] Riksförsäkringsverket (2000) *Sjukfrånvaro och förtidspension,* RFV ANALYSERAR 2000:2. Stockholm: Riksförsäkringsverket (in Swedish).

[17] SOU (2002) (In Swedish) *Handlingsplan för ökad hälsa i arbetslivet.* Statens Offentliga Utredningar, 2002:2, Stockholm: Fritzes.

APPENDIX

Predictor	Label	No. of missing values (%)
Z_1	Age $\begin{cases}1 \text{ if } \geq 31 \text{ years} \\ 0 \text{ if } < 31 \text{ years}\end{cases}$	0(0)
Z_2	Education $\begin{cases}1 \text{ if low} \\ 0 \text{ if high}\end{cases}$	208(38)
Z_3	Ethnicity $\begin{cases}1 \text{ if Non-swedish} \\ 0 \text{ if Swedish}\end{cases}$	170(31)
Z_4	Household income $\begin{cases}1 \text{ if } \geq 7000 \text{ SEK} \\ 0 \text{ if } < 7000 \text{ SEK}\end{cases}$	239(44)
Z_5	Demand $\begin{cases}1 \text{ if high} \\ 0 \text{ if low}\end{cases}$	267(49)
Z_6	Control $\begin{cases}1 \text{ if low} \\ 0 \text{ if high}\end{cases}$	267(49)
Z_7	Strain $\begin{cases}1 \text{ if high} \\ 0 \text{ if low}\end{cases}$	267(49)
Z_8	Attitude $\begin{cases}1 \text{ if high moral} \\ 0 \text{ if low moral}\end{cases}$	180(33)
Z_9	Inconvenient working environment $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	263(48)
Z_{10}	Heavy lifts $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	265(49)
Z_{11}	Suitable working tasks $\begin{cases}1 \text{ if no} \\ 0 \text{ if yes}\end{cases}$	333(61)
Z_{12}	Sick-listing in the family $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	187(34)
Z_{13}	TDP/ER in the family $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	181(33)
Z_{14}	Offered TDP/ER $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	260(48)
Z_{15}	Work ability $\begin{cases}1 \text{ if } \text{bad} (\leq 4) \\ 0 \text{ if } \text{good} (> 4)\end{cases}$	178(33)
Z_{16}	Comorbidity $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	173(32)
Z_{17}	Smoking $\begin{cases}1 \text{ if yes or never} \\ 0 \text{ if quited}\end{cases}$	184(34)
Z_{18}	Rehabilitation plan $\begin{cases}1 \text{ if yes} \\ 0 \text{ if no}\end{cases}$	7(1)

Table A1: Labels to predictors and the number of missing values (%) for MB$(n=545)$. The abbreviation TDP/ER denotes Temporary Disability Pension/Early Retirement.

Z_1	Z_3	Z_5	Z_{11}	Z_{12}	Z_{15}	Z_{16}	Z_{18}	$\hat{\pi}$	$\hat{\pi}_{Lower}$	$\hat{\pi}_{Upper}$	$\hat{V}[\hat{\pi}]$
1	1	1	0	0	0	1	0	0.03	0.01	0.61	0.001
1	0	1	0	0	0	0	0	0.08	0.04	0.24	0.001
0	0	1	0	0	0	0	0	0.08	0.04	0.38	0.002
1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.14**	**0.09**	**0.26**	**0.001**
0	0	0	0	0	0	0	0	0.14	0.07	0.41	0.005
0	0	0	0	1	1	0	1	0.18	0.06	0.68	0.029
1	0	0	1	0	0	0	0	0.19	0.11	0.38	0.004
0	0	0	1	0	0	0	0	0.20	0.09	0.50	0.010
1	1	1	0	0	0	0	0	0.23	0.09	0.61	0.021
1	0	1	1	0	0	0	0	0.23	0.15	0.40	0.004
1	1	1	1	0	0	0	0	0.30	0.15	0.58	0.015
0	0	1	0	0	1	0	0	0.30	0.14	0.61	0.020
1	0	1	0	0	1	0	0	0.31	0.18	0.54	0.010
1	1	1	1	1	1	1	0	0.37	0.12	0.78	0.080
1	1	0	0	0	0	0	0	0.39	0.19	0.67	0.024
1	0	1	0	1	0	0	0	0.41	0.19	0.71	0.033
1	1	1	1	0	1	1	0	0.44	0.19	0.74	0.040
0	0	0	0	0	1	0	0	0.46	0.25	0.70	0.021
0	0	0	1	1	1	0	1	0.46	0.16	0.81	0.084
1	0	1	0	0	1	0	1	0.47	0.19	0.78	0.051
1	0	0	0	0	1	0	0	0.47	0.33	0.62	0.007
0	0	1	1	1	1	0	0	0.51	0.22	0.79	0.045
0	0	0	1	0	1	0	0	0.55	0.29	0.77	0.025
1	0	1	0	1	1	0	0	0.55	0.27	0.79	0.034
1	0	0	1	0	1	0	0	0.56	0.37	0.72	0.011
1	0	0	0	0	1	0	1	0.57	0.29	0.79	0.030
1	0	0	0	1	0	0	0	0.57	0.30	0.79	0.027
0	1	1	0	0	1	0	0	0.60	0.24	0.85	0.058
0	0	1	1	0	1	0	0	0.61	0.34	0.80	0.021
1	0	1	1	0	1	0	0	0.62	0.43	0.75	0.008
1	0	0	1	1	0	0	0	0.66	0.34	0.85	0.027
1	0	1	0	0	0	0	1	0.66	0.22	0.89	0.082
1	1	1	1	0	1	0	0	0.69	0.41	0.85	0.016
1	0	0	0	1	1	0	0	0.71	0.40	0.86	0.019
1	0	1	1	1	0	0	0	0.71	0.38	0.87	0.021
1	0	0	0	0	0	0	1	0.74	0.29	0.92	0.046
1	1	1	0	0	1	0	1	0.75	0.28	0.92	0.049
1	0	0	1	0	1	1	0	0.76	0.30	0.92	0.040
0	0	0	1	0	1	1	0	0.76	0.26	0.93	0.060
1	1	0	0	0	1	0	0	0.77	0.45	0.89	0.014
1	0	0	1	1	1	0	0	0.78	0.43	0.90	0.016
1	0	1	1	1	1	0	0	0.81	0.47	0.92	0.011
1	0	1	1	0	1	1	0	0.83	0.35	0.94	0.024
1	1	1	1	0	1	1	1	0.83	0.37	0.94	0.021
1	0	0	1	0	1	0	1	0.84	0.50	0.93	0.009
1	0	1	1	0	1	0	1	0.91	0.62	0.96	0.003
1	0	0	1	0	0	0	1	0.92	0.39	0.97	0.007
1	1	1	1	0	1	0	1	0.93	0.54	0.97	0.003
1	0	1	1	0	0	0	1	0.96	0.43	0.99	0.002
1	**0**	**1**	**1**	**0**	**1**	**1**	**1**	**0.97**	**0.38**	**0.99**	**0.002**

Table A2: Ordered predicted values and associated CI's for various combinations of **Z** in accordance with Figure 1 (MB90). The variance corresponds to formula (8) in Jonsson and Persson (2002). See also Table A1 for labels to the predictors.

Legends To Figures

Figure 1: The probability of 'no work resumption' and associated CI's (MB90).

Figure 2: The probability of 'no work resumption' and associated CI's (MB1Y).

Figure 3: The probability of 'no work resumption' and associated CI's (MN90).

Figure 4: The probability of 'no work resumption' and associated CI's (MN1Y).

Figure 5: The probability of 'no work resumption' and associated CI's (WB90).

Figure 6: The probability of 'no work resumption' and associated CI's (WB1Y).

Figure 7: The probability of 'no work resumption' and associated CI's (WN90).

Figure 8: The probability of 'no work resumption' and associated CI's (WN1Y).

Figure 9: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Age (MB90).

Figure 10: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Rehabilitation plan (MB90).

Figure 11: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Comorbidity (MB90).

Figure 12: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Work ability (MB90).

Figure 13: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Demand (MB90).

Figure 14: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Sick-listing in the family (MB90).

Figure 15: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Suitable working tasks (MB90).

Figure 16: Estimated differences $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_0$ for Ethnicity (MB90).

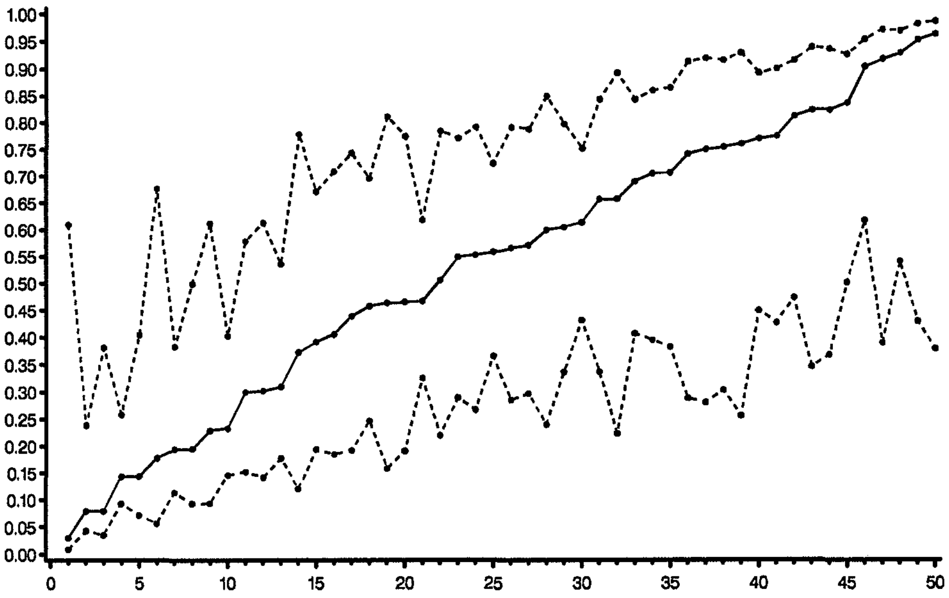


Figure 1

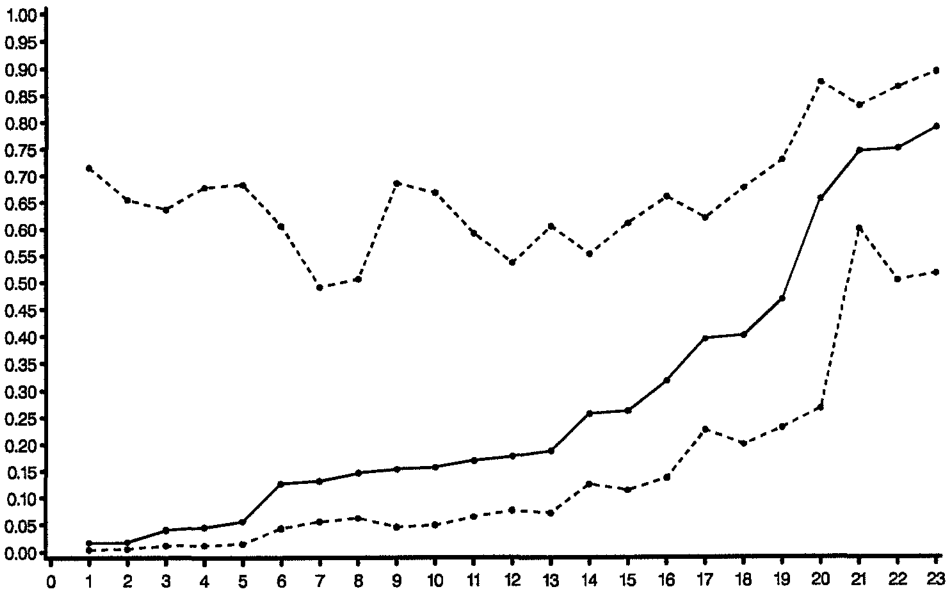


Figure 2

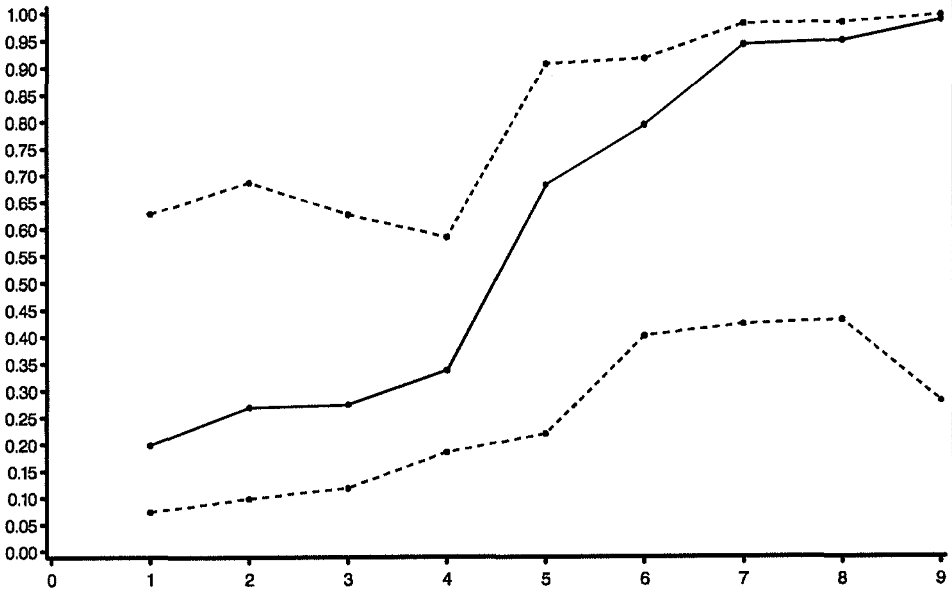


Figure 3

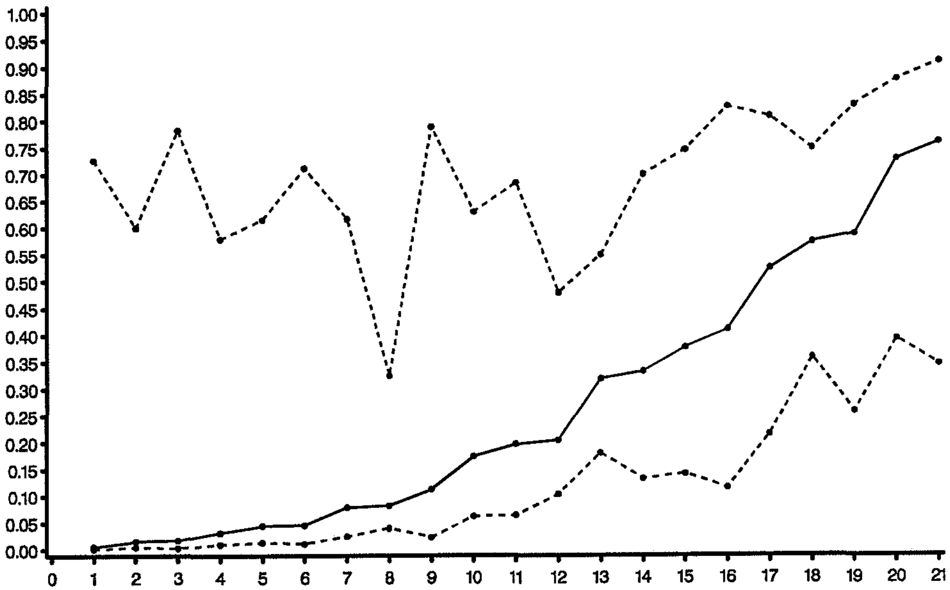


Figure 4

Figure 5

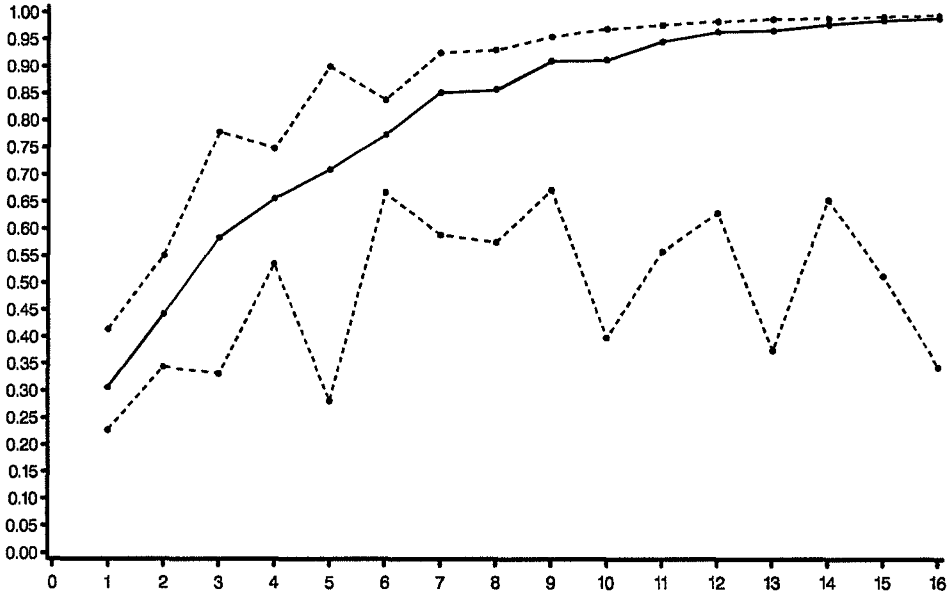


Figure 6

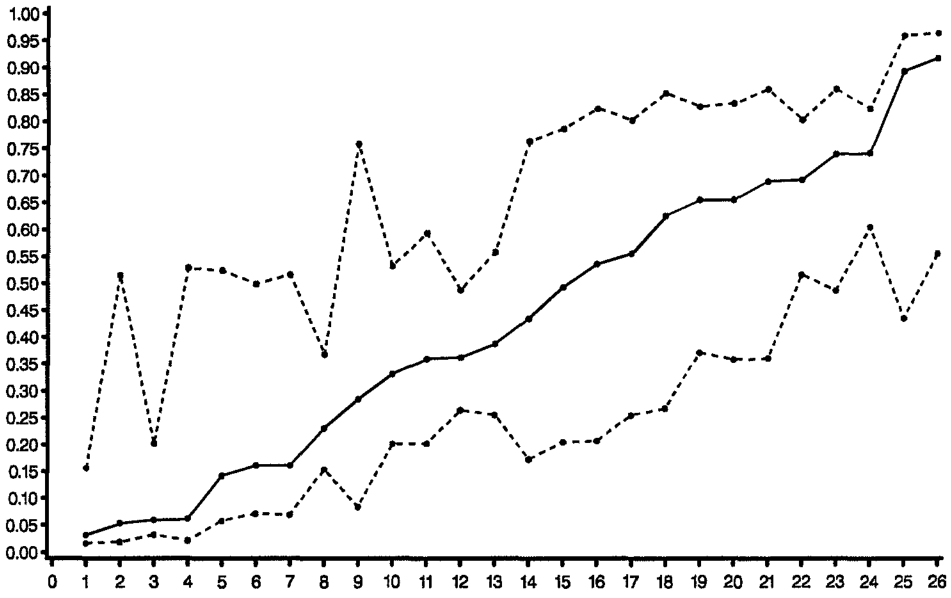


Figure 7

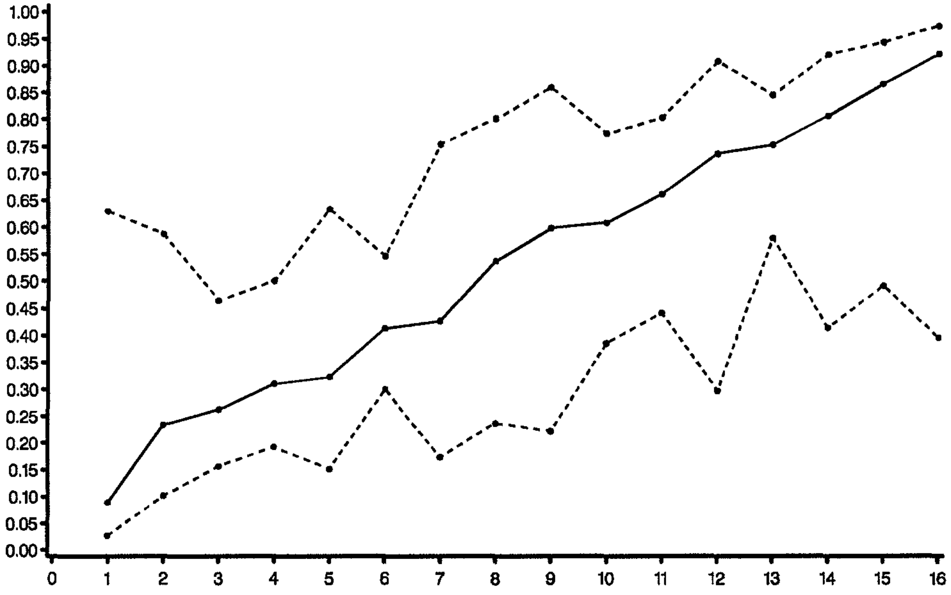


Figure 8

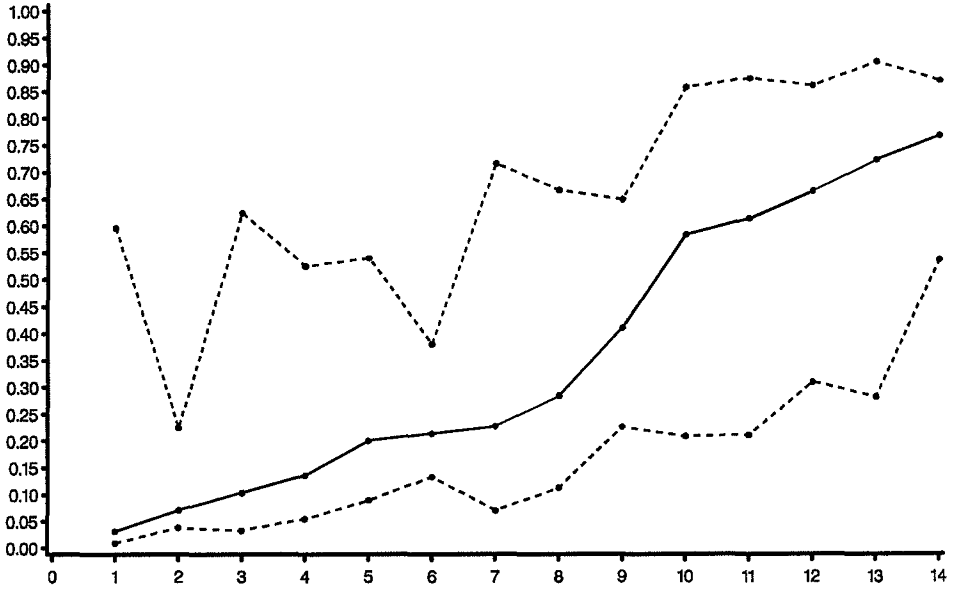


Figure 9

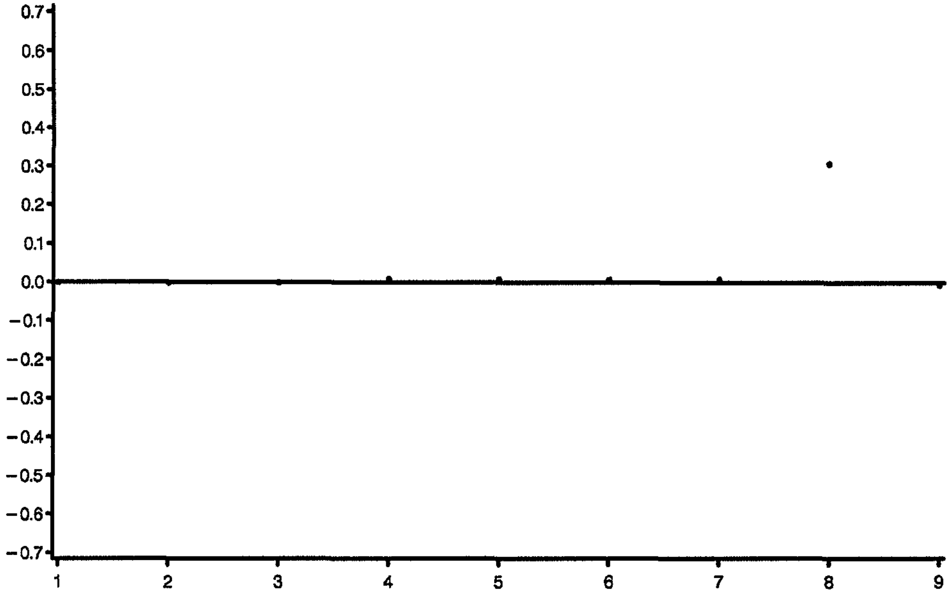


Figure 10

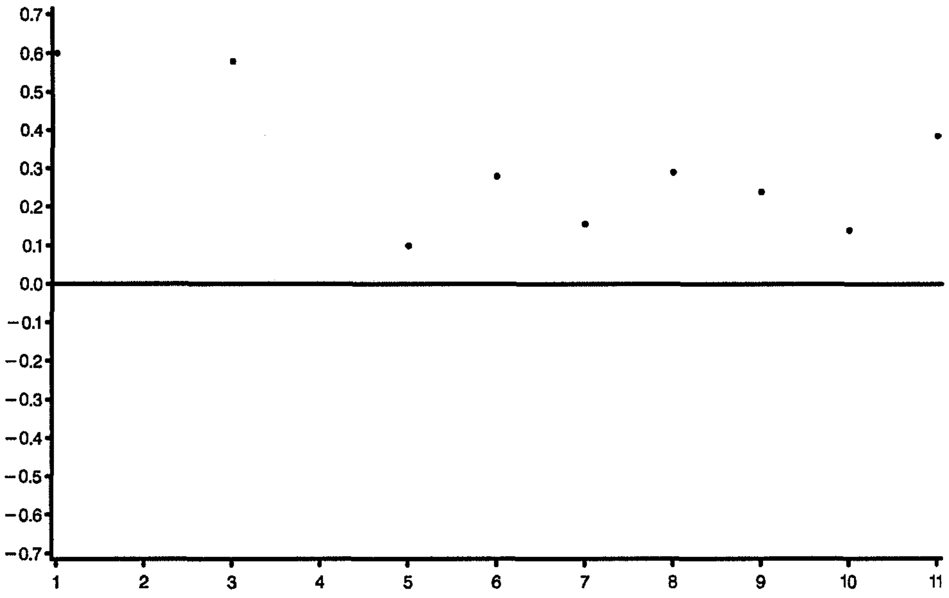


Figure 11

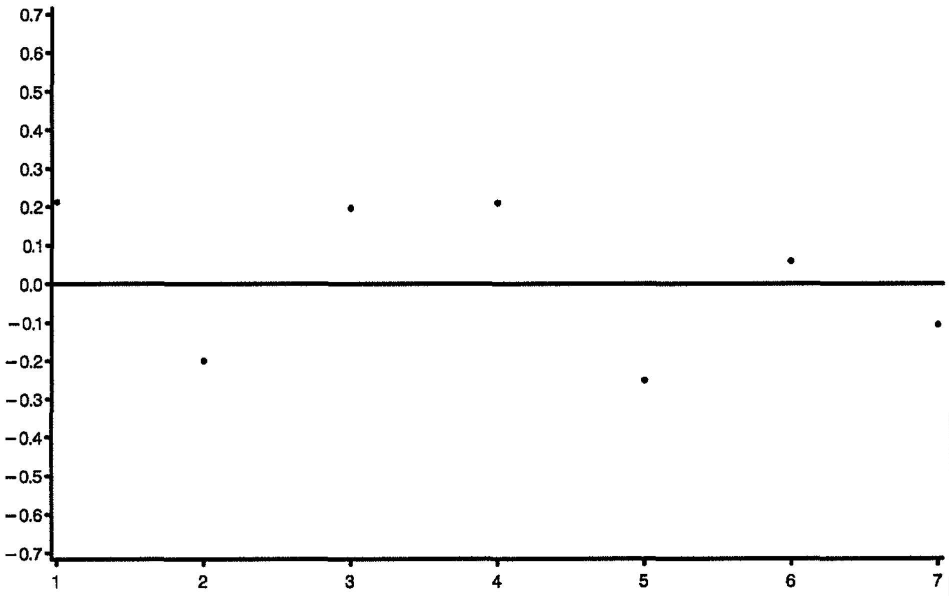


Figure 12

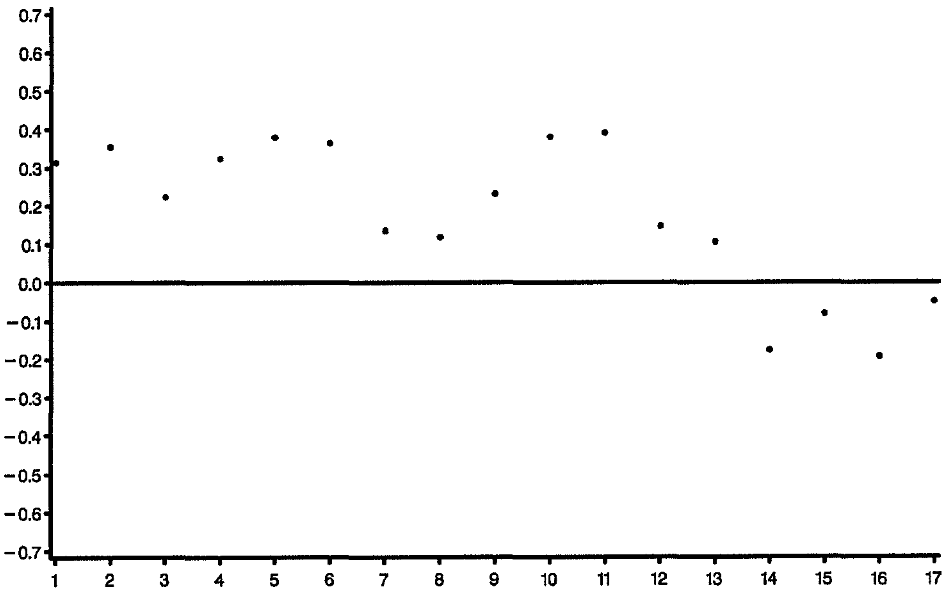


Figure 13

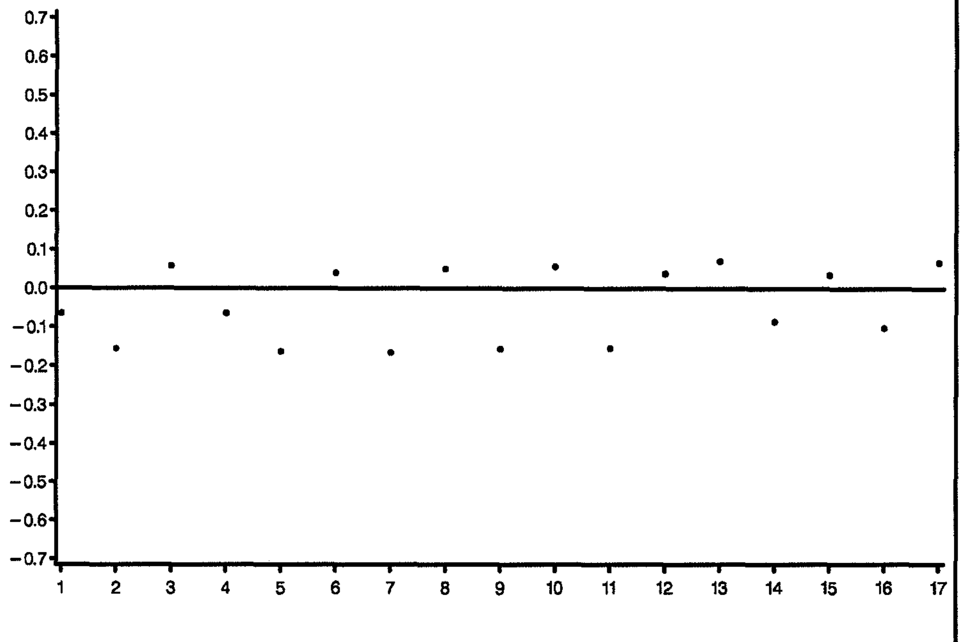


Figure 14

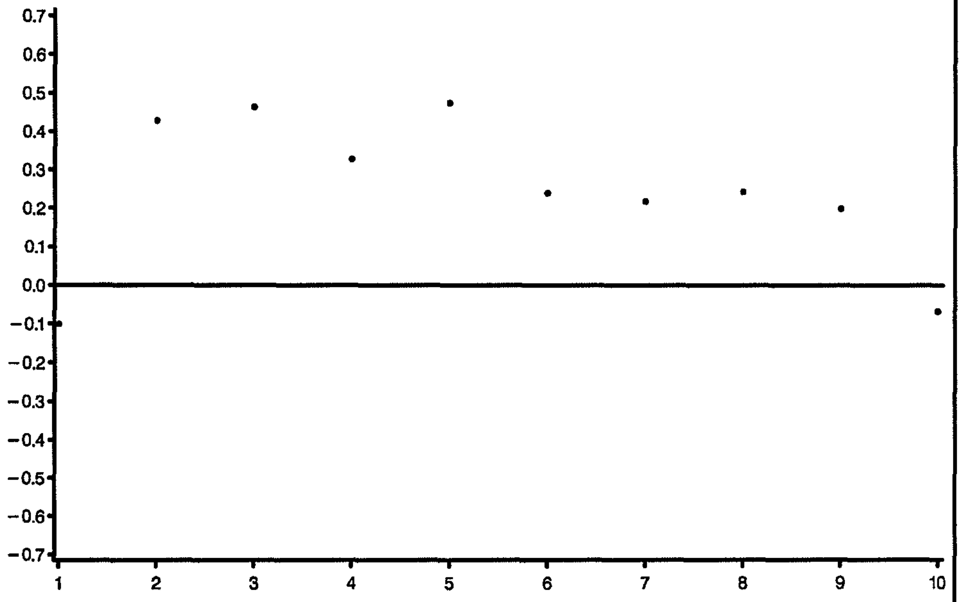


Figure 15

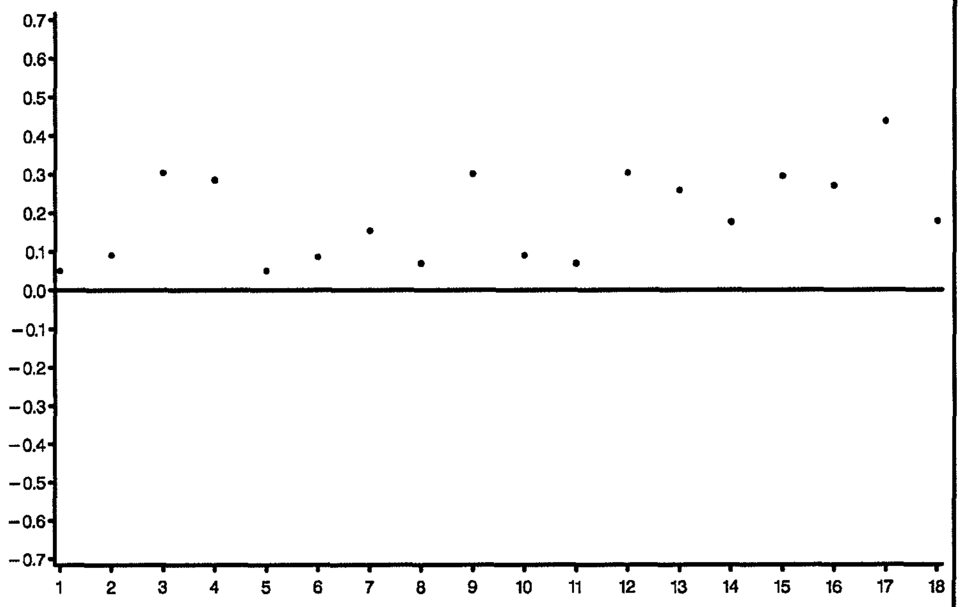
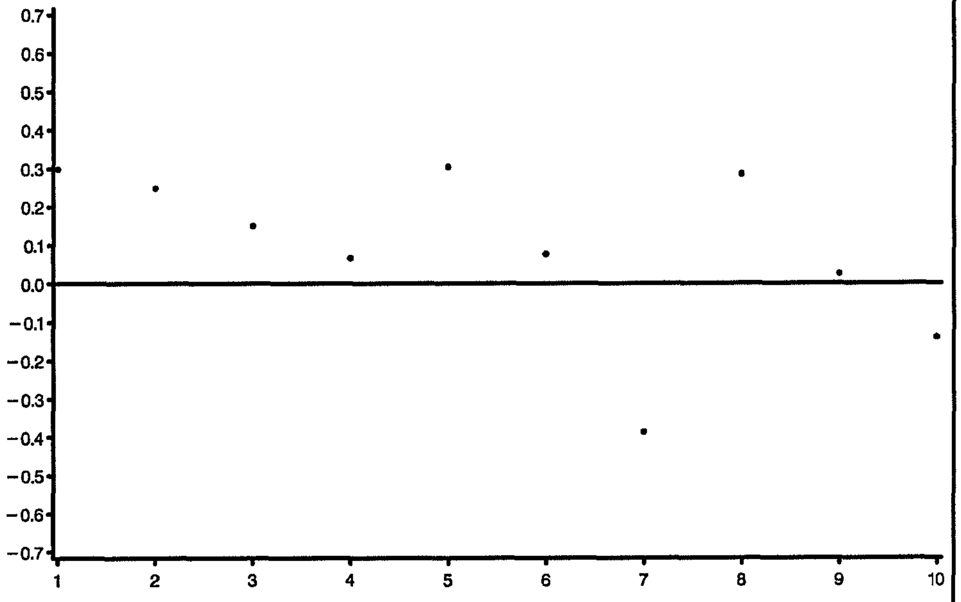


Figure 16



Research Report

2001:1	Holgersson, H.E.T.:	On assessing multivariate normality.
2001:2	Sonesson, C. & Bock, D.:	Statistical issues in public health monitoring – A review and discussion.
2001:3	Andersson, E.:	Turning point detection using non-parametric statistical surveillance. Evaluation of some influential factors.
2001:4	Andersson, E. & Bock, D.:	On seasonal filters and monotonicity.
2001:5	Andersson, E., Bock, D. & Frisén, M.:	Likelihood based methods for detection of turning points in business cycles. A comparative study.
2001:6	Sonesson, C.:	Evaluations of some exponentially weighted moving average methods.
2001:7	Sonesson, C.:	Statistical surveillance. Exponentially weighted moving average methods and public health monitoring.
2002:1	Frisén, M. & Sonesson, C.:	Optimal surveillance based on exponentially weighted moving averages.
2002:2	Frisén, M.:	Statistical surveillance. Optimality and methods.
2002:3	Jonsson, R. & Persson, A.:	Bayes prediction of binary outcomes based on correlated discrete predictors.
2002:4	Persson, A.:	Prediction of work resumption among men and women with lower back- and neck pain in a Swedish population.