On the use of Expected Attainable Discrimination for feature selection in large scale medical risk prediction problems

D. R. Lovell, M. J. J. Scott, M. Niranjan,
R. W. Prager, K. J. Dalton and R. Derom

CUED/F-INFENG/TR299

September 12, 1997

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
United Kingdom

On the use of Expected Attainable Discrimination for feature selection in large scale medical risk prediction problems

D. R. Lovell, M. J. J. Scott, M. Niranjan, R. W. Prager, K. J. Dalton and R. Derom*
{drl,mjjs,niranjan,rwp}@eng.cam.ac.uk,
kjd5@cam.ac.uk

September 12, 1997

Abstract

This report investigates the use of *expected attainable discrimination* (EAD) as a measure to select discrete valued features in two-class prediction problems. In essence, EAD tells us the performance we could expect to achieve with a simple histogram probability density model of a given dataset. For discrete valued features, this kind of density model is *bias-free* but can have large *variance*. Given insufficient training data, such a model's test set performance will be lower than that of a suitably biased model. In light of this, we explore the usefulness of EAD for feature selection.

Keywords: Feature selection, area under receiver operating characteristic (ROC) curve, medical risk prediction, obstetrics.

1 Introduction

Feature selection is the process of choosing a good representation of data to solve a given inference problem. We are interested in choosing a representation of a mother's health state that will allow us to accurately model obstetrical risk. This means we have to decide which features (also referred to as *predictors*, or simply *variables*) should be used to forecast a mother's risk of experiencing a particular adverse pregnancy outcome (APO).

When the Quality Assurance in Maternity Care (QAMC) project commenced in 1995, we were faced with the problem of determining good predictors of APOs from a large set of candidate variables. In contrast to many medical risk prediction scenarios, we had almost *too much* data. Within the Scottish Morbidity Record SMR2, for example, there are over 770 000 records. Each record can contain up to 6 different ICD-9 codes [1] to describe maternal condition. Each ICD-9 code can represent one of 300 possible maternal conditions arising prior to the onset of labour. The methods discussed in this paper were developed as a practical means to tackle feature selection in this setting.

There are two main kinds of feature selection methods: those based on probabilistic separability measures applied to the data, and those based on the error rate of a classifier as a design criterion. This report focuses on the latter approach. As Siedlecki and Sklansky [2, Section 3.4] point out, such an approach "...must be recognized as a process in which [the] classifier is optimized and, therefore, trained. Hence, the only error rate that can be computed for this classifier is an apparent error rate, which is known to be very biased". The feature selection criterion proposed in this paper is based around a classification model that is both fast to train and, through the use of hold-out data, able to provide a performance estimate less biased than apparent error rate.

Regardless of the bias of this performance estimate, it may still under or over-estimate the performance that could be obtained *by using a different model* to classify the data. Consequently, our feature selection

*The first four authors may be contacted at the Cambridge University Engineering Department, Trumpington St., Cambridge CB2 1PZ, UK. K. J. Dalton is with the Cambridge University Department of Obstetrics and Gynaecology, Rosie Maternity Hospital, Cambridge CB2 2SW, UK. R. Derom is with the Katholieke Universiteit Leuven, Center for Human Genetics, Herestraat 49, B-3000 Leuven, Belgium.

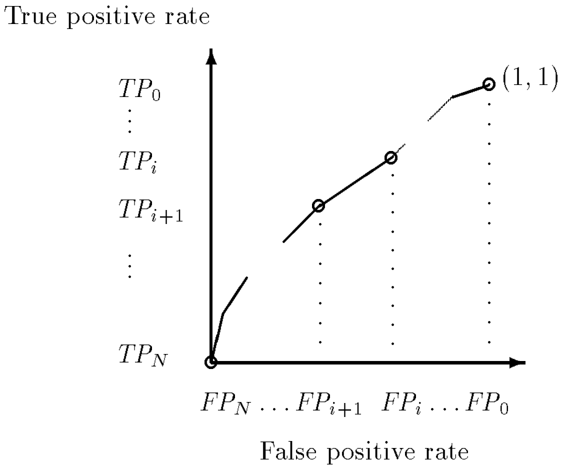

Figure 1: These sections of an ROC curve show true and false positive rates across the thresholds, $t_0, \ldots, t_N$, separating bins, $B_1, \ldots, B_N$, in an ordered list. Straightforward methods to calculate the area under this curve (and its standard error) are presented in [3] and [4].

method may select too few or too many features for other models to accurately classify data. The aim of this report is to explore this issue and determine the strengths and weaknesses of this approach to selecting discrete valued features in large databases.

In Section 2, we present the concept of *expected attainable discrimination* (EAD) and show how it can used in a sub-optimal search for discriminative subsets of features. Section 3 gives performance results obtained with a variety of risk prediction models that use features selected by EAD. This report concludes with a discussion of the problems in assessing feature selection methods and a statement of the advantages and shortcomings of EAD as a feature selection criterion.

2 Feature selection using EAD

Our ultimate goal is to forecast the risk associated with a mother's health state, *not* to classify the state as "at risk" or "not at risk". While a decision about whether to intervene must be made eventually, it is not appropriate for us to make that decision for the clinician managing the case. So, unlike many pattern recognition problems, we do not focus on classification/misclassification rates but speak instead of discrimination between adverse and benign outcomes. As we shall see in Section 2.2, discrimination is maximized by a system that predicts the actual risk of adverse outcome[1] associated with a given health state.

2.1 Problem setting

We wish to choose features that will allow us to discriminate between patterns belonging to class $\mathcal{P}$ and patterns belonging to class $\mathcal{Q}$. Each training pattern is represented by a vector of discrete valued features, $\boldsymbol{x} = (x_1, \ldots, x_n)$, and a class label, $y \in \{\mathcal{P}, \mathcal{Q}\}$. We implement a *histogram probability density model* by grouping patterns with identical features into *bins* such that bin B_i contains p_i patterns from class $\mathcal{P}$ and q_i patterns from class $\mathcal{Q}$. With each bin, we associate an estimate of the probability that data from that bin belong to class $\mathcal{P}$, that is, $\hat{P}(y = \mathcal{P} | \boldsymbol{x} \in B_i)$ for bin i. This imposes an ordering on the bins that we

[1] or some monotonically increasing function thereof.

can represent as:

$$\begin{array}{ccccccccc} & B_1 & & B_2 & & & & B_N & \\ & \boxed{\begin{array}{c}p_1\\q_1\end{array}} & & \boxed{\begin{array}{c}p_2\\q_2\end{array}} & & \cdots & & \boxed{\begin{array}{c}p_N\\q_N\end{array}} & \\ \uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ t_0 & & t_1 & & t_2 & & t_{N-1} & & t_N \end{array}$$

where bins are arranged from left to right in ascending order of $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$, and classification thresholds $t_0, \ldots, t_N$ separate each bin in the list. Using this representation, we define

$$\text{true positive rate at } t_i\text{:} \quad TP_i = \sum_{j>i}^{N} \frac{p_j}{|\mathcal{P}|} \tag{1}$$

$$\text{false positive rate at } t_i\text{:} \quad FP_i = \sum_{j>i}^{N} \frac{q_j}{|\mathcal{Q}|}, \tag{2}$$

where $|\mathcal{P}|$ and $|\mathcal{Q}|$ denote the number of patterns in classes $\mathcal{P}$ and $\mathcal{Q}$, respectively. The *receiver operating characteristic* (ROC) curve (Figure 1) plots TP_i versus FP_i across all thresholds t_i. Furthermore, the *area under the ROC curve* (AUROC) measures how well a prediction system discriminates between two classes of outcome [5].

The AUROC depends on how bins are ordered, and that is determined by the estimates of $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$. In Section 2.2 we discuss the form of these estimates that maximizes the AUROC. Many classifiers work by estimating $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$ (or some monotonically increasing function thereof) and then applying a threshold to that estimate, above which $\boldsymbol{x}$ is deemed to belong to class $\mathcal{P}$. The value of the threshold is determined by the misclassification costs and the class prior probabilities. The AUROC provides a summary of the classifier's behaviour across *all* thresholds and is, thus, an appropriate measure of performance when misclassification costs are not clearly specified.

Unlike most measures of probabilistic separation, the AUROC has a number of physical interpretations, depending on the experimental setting in which it was obtained [5, Chapter 2]. In this paper, the following interpretation is relevant: given two patterns, one randomly chosen from class $\mathcal{P}$, the other randomly chosen from class $\mathcal{Q}$, the probability that a classifier correctly decides which pattern is class $\mathcal{P}$ is equal to the area under the classifier's ROC curve. Throughout the remainder of this paper, we shall use the terms "discrimination" and "area under the ROC curve" synonymously.

2.2 Maximum attainable discrimination

It is possible to calculate the maximum extent to which one can discriminate between two classes within a given set of data. From Equations 1 and 2 we have that

$$\begin{aligned} TP_i &= TP_{i-1} - \frac{p_i}{|\mathcal{P}|} \\ FP_i &= FP_{i-1} - \frac{q_i}{|\mathcal{Q}|}. \end{aligned}$$

Since $|\mathcal{P}|$ and $|\mathcal{Q}|$ are constant in a given data set, the slope of the i^{th} segment of the ROC curve (between adjacent points $(FP_i, TP_i), (FP_{i-1}, TP_{i-1})$) is proportional to p_i/q_i, and equal to the estimated likelihood ratio in the i^{th} bin. Thus, if we place bins in ascending order of estimated likelihood ratio[2], the line segments that make up the ROC curve appear, from left to right, in decreasing order of slope, and the AUROC is maximized (see step (1) in Figure 2) [5, 7, 8].

This *maximum attainable discrimination* — or *MAD*, as Hanley and McNeil [8] refer to it — is easily calculated and applies to *any* system which associates outputs with discrete valued data in a consistent manner (*i.e.*, labels identical patterns identically). As we shall see in Section 3.3, MAD helps determine

[2] or, equivalently, $p_i/(p_i + q_i)$. This maximum likelihood estimate of the probability that data from bin B_i belongs to class $\mathcal{P}$ is what we would obtain using a simple histogram model [6] of the probability density function underlying the data set.

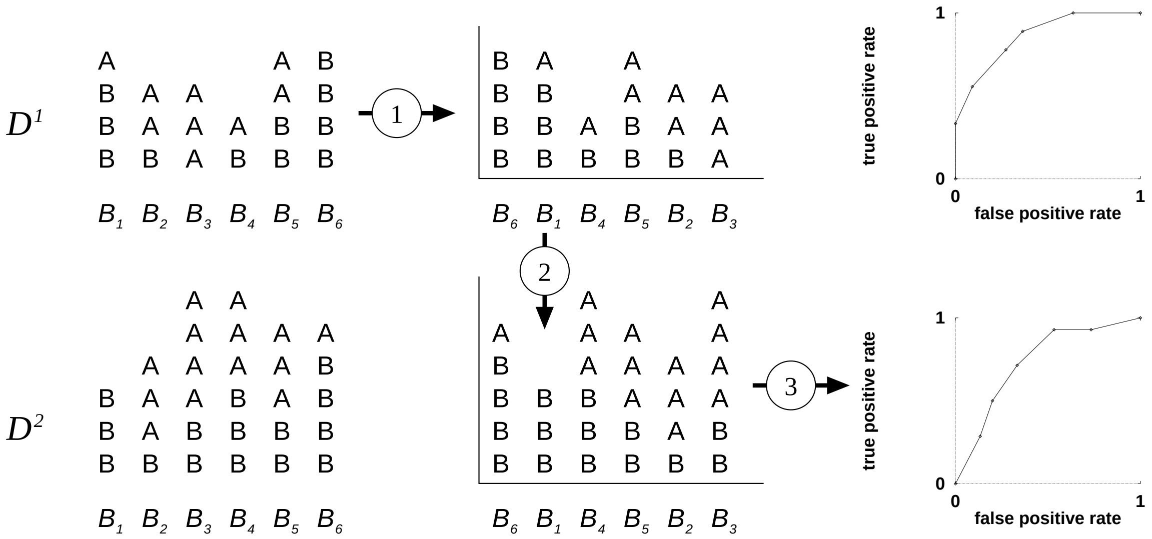

Figure 2: To estimate the discrimination we can *expect* to obtain with a given representation of a specific dataset, we randomly partition our dataset into two disjoint subsets, D^1 and D^2. The D^1 bins, B_1–B_6, are placed in MAD order (1). The D^1 MAD ordering is applied to the bins in the D^2 subset (2). The area under the D^2 ROC curve is an estimate of the expected discrimination (3).

whether a prediction system's performance is limited by an inherent lack of separability in the data, or shortcomings of the prediction model.

Note that MAD increases with the number of features used to represent a given data set. However, our ability to generalize from the characteristics of a finite amount of data tends to *decrease* once a certain representational complexity is exceeded.

2.3 Expected attainable discrimination

One aim of feature selection is to choose from a set of available predictors, those which afford the greatest degree of discrimination between two classes of data. Implicit in this objective are the ideas that a prediction system will eventually be built using the selected representation(s) of the data, and that that system will be used to make predictions about *new* data.

A natural way to select features is simply to evaluate the performance of a prediction model using a variety of representations of new data. Ignoring the combinatorial nature of the number of possible feature subsets for the time being, we propose a practical classification scheme and a way to estimate its performance on new data.

The scheme uses the histogram probability density model described in Section 2.1 to estimate *expected attainable discrimination* (EAD). Given two sets of data, D^1 and D^2, drawn from the population we wish to classify (see Figure 2), we

1. obtain the MAD with D^1 by placing its bins in ascending order of $p_i/(p_i + q_i)$ to form the list, L^1;
2. apply that ordering to the corresponding bins in D^2 to give L^2, and measure the discrimination obtained with that list.

The area under the ROC curve for L^2 estimates EAD, and indicates how well the features representing the data allow us to predict the outcome of new data.

Any bins that appear in D^2 but not D^1 are inserted randomly into the L^2 ordering. There are other ways to deal with these "unassigned" bins such as the Bayesian approach described by Peto and MacKay [9], however this method was used throughout the experiments described in this paper. (The basic principle of Peto and MacKay's approach is described in Appendix C.)

Data sets D^1 and D^2 should be independent to avoid highly biased estimates of the prediction system's performance. Methods such as cross-validation and the bootstrap exist to generate training and test data sets from a single data set so as to mitigate such bias (see [10] for a summary). Given the large amount of data available to us, in our experiments, it was convenient to randomly select two-thirds of our data to create set D^1 and use the remaining third to form D^2. A *rotation error* estimate [2, p. 216] of EAD was formed by averaging EAD across ten such random partitions.

2.4 Using EAD to select discriminative features

EAD provides a measure of the degree of discrimination we can expect to achieve when a particular subset of features is used to represent a given data set. Thus we can use EAD to search for the best subset of features among a number of feasible subsets. Since this search is NP hard, we must resort to sub-optimal methods [11].

The experiments described in this paper use *sequential forward selection* with EAD (which we refer to as FEAD). (Results of the more computationally intensive sequential backwards selection (BEAD) are reported elsewhere [12].) FEAD is an iterative search method defined by

$$W_{n+1} \quad = \quad W_n + \underset{x_i \notin W_n}{\operatorname{argmax}} \ \mathrm{EAD}(W_n + \{x_i\}), \tag{3}$$

where the working set W_n contains the n discrete features used to describe the data at the current search step. Using the methods described in the previous section, $\mathrm{EAD}(W_n)$ returns an estimate of the discrimination we can expect when the features in W_n are used to represent a given data set. Thus, features will be added from the working set in order of decreasing importance to attainable discrimination. Since $\mathrm{EAD}(W_n)$ is an estimate based on several random partitions of the data, the order in which features are selected is not deterministic. For this reason, we recommend performing several runs of the algorithm to observe *trends* in the order of feature selection.

The fact that different runs of FEAD or BEAD can produce different orderings of features may trouble some readers at first. It is tempting to expect there to be some combination of factors that consistently *explain* an outcome, especially a clinical outcome. However there is no guarantee that our data contains any observations of those factors, nor is there a guarantee that we could determine such factors — had they been recorded — using a finite amount of data. As we shall see in the next section, experiments show that strong predictors of outcome are consistently selected using EAD. The selection of features less strongly associated with the outcome is more influenced by the random partitioning of the data.

Before presenting empirical results obtained with FEAD, we remind the reader that the novel aspect of this research is in the use of EAD as a measure of the discriminatory power of a given subset of features — not in the use of any particular search technique. We have elected to use sequential selection as a search heuristic because it is an uncomplicated and fairly successful strategy. The performance of *floating search methods* [13] makes them an appealing alternative.

3 Experimental application of EAD

In Section 2 we described how EAD can be used in selecting features. EAD is an average of the AUROCs obtained when we run test data through several simple histogram density models. Each model is built using a random training/test split of the available data. Thus, FEAD chooses subsets of features which allow simple histogram density models to perform well on the available data. We can then use those subsets of features in the construction of other models, such as neural networks.

But why use the histogram density model at all? Why not search through the subsets of features using, say, neural network performance as a selection criterion? Two reasons for using the histogram density model are that it is *fast* to train and, for discrete valued data, free from *bias*, *i.e.*, it has the necessary power to represent *any* mapping from the space of discrete valued inputs to a real valued output.

There is a catch. Bias-free models tend to have large *variance* [14] and require much more training data than their biased cousins to learn to generalize effectively. The number of parameters in a histogram density model can increase exponentially with the number of features used to represent the data. Consequently, as feature subset size increases, EAD tends to become more pessimistic than the performance achievable by other, less flexible models.

So can these pessimistic performance estimates be used in the selection of discriminative features? In particular, do the features selected using EAD give rise to good performance in other models? To explore this issue, we look at the performance of a variety of models built using different subsets of features. Before we describe these models, we introduce the practical prediction task used in this investigation.

3.1 Prediction of failure to progress in labour

One outcome considered in the QAMC project is *Cæsarean section for failure to progress in labour*. We consider the precursor to that outcome: failure to progress in labour — a condition which means that the natural course of labour has stalled before delivery, placing the wellbeing of both mother and child at risk.

Our experiments were based on data from the Scottish Morbidity Record (known as the SMR2): 771571 singleton births that occurred between 1980–91. Cases in which elective Cæsarean section or breech presentation occurred were removed from the dataset[3]. These records were partitioned into a *learning set* of births occurring between 1980–88 (540905), and a *testing set* of births occurring between 1989-91 (176812). While we only have access to *retrospective* data, this partition is used to simulate a *prospective* trial of the prediction systems.

Implicit in this use of the SMR2 data is the assumption that the characteristics of the population remain stationary. Investigation of changes in the risk of adverse outcome over time is certainly a worthwhile objective but it is beyond the scope of this report.

3.2 Selection of predictor variables

The SMR2 data contains many features relating to conditions prior to the onset of labour (over 300 distinct ICD-9 codes appear). From this set of features, a subset of 49 features was pre-selected using Cramer's V statistic [15] as a univariate measure of the strength of association between each feature and the outcome. While not strictly necessary, this pre-selection relieved much of the computational burden on the FEAD algorithm. We acknowledge that this step could possibly remove variables important to the *multivariate* prediction of risk, but our aim here is to illustrate the use of EAD for feature subset selection, not to come up with a clinically unassailable prediction model.

Using the 49 candidate features, 20 separate runs of FEAD were performed. Figure 3 shows the MAD and EAD obtained at each forward selection step of the 20 runs. Clearly, once a certain representational complexity is reached, using additional features to describe the data results in a *decrease* in EAD. Figure 4 shows the levels of EAD up to point of maximum EAD in each run. Table 1 lists the order in which features were selected in 10 of the 20 runs performed, and shows in bold the features in the subsets which attained maximum EAD.

Although the different runs of FEAD give an indication of the discriminative power of each predictor, we must still choose which variables to build prediction models with. A number of approaches spring to mind:

1. Select those predictors common to *all* subsets of features that attained maximum EAD in *all* runs of FEAD. If we considered only 10 runs of FEAD, this approach would select the predictors in the first eight rows of Table 1.

2. Select those predictors that were in *any* of the subsets of features that attained maximum EAD in *any* run of FEAD. If we considered only 10 runs of FEAD, this approach would select predictors with a bold face entry anywhere in their row of Table 1.

3. Build an *ensemble* of prediction models, each based on a subset of features that attained maximum EAD on one run of FEAD.

4. Ignore the results of FEAD and select predictors according to some other criteria.

5. Ignore the results of FEAD and use all available predictors.

[3] In almost all cases, these factors will be known before labour commences and indicate that delivery will be by Cæsarean section so that failure to progress in labour cannot occur.

We did not pursue the first approach since we knew that a model based on the those eight predictors would be outperformed by a simple histogram density model[4].

The second approach resulted the set of 35 variables listed in Table 3. The third approach meant that 20 different subsets of variables were used to construct the individual members of an *ensemble* of neural networks (see the description of the NetMix model in Appendix B). Note that both the second and third approaches ignore the variance in our estimates of EAD. We do not address the question of how to find the subset of variables that attains the largest value of EAD that is *significantly different* from the EAD of other subsets.

As for the fourth approach, after discussions with clinicians, it became clear that there were concerns about the reliability and validity of the ICD-9 codes used to describe maternal condition. In response to these concerns, we decided to explore the performance of prediction systems trained solely on non-ICD code information. This criterion resulted in data described by the 4 variables `Parity`, `MumAge`, `CSections` and `NeoDxs`.

So, at the end of the feature selection process we had decided to investigate 4-, 35-, and 49-variable feature sets, as well as the mixture of 9–15-variable feature sets to be used in the NetMix model.

Representation of information is an important issue in modelling. While the majority of variables were already binary valued, some variables, *e.g.*, `MumAge`, were polytomous (see Table 4). These variables were encoded using the *reference variable* method described by Hosmer and Lemeshow [16].

3.3 Prediction performance with selected features

The central purpose of this report is to assess whether features selected using a cheap and cheerful heuristic produce good performance in sophisticated prediction models that are more time consuming to train. In this section, we describe the prediction performances achieved with four different prediction models: logistic regression (LogReg), a smoothed lookup table (LkpTab), an ensemble of neural networks with the same architecture (NetEns), and an ensemble of neural networks with different architectures (NetMix). (These models are described in Appendices A, B and C.)

Figure 5 shows the performance of the each prediction system with 4-, 35-, and 49-variable feature sets, as well as the mixture of 9–15-variable feature sets used in the NetMix model. These results suggest that the underlying structure of the association between predictors and outcome can be captured with fewer than 49 features. The AUROC and *average negative log-likelihood* (ANLL) of the test data for each prediction model are given in Table 2. ANLL indicates how accurately a model predicts the probability of adverse outcome observed in the test data:

$$\begin{aligned} \text{ANLL} &= -\frac{1}{N}\sum_{\boldsymbol{x}_i \in X} p_i \log P(\text{adverse}|\boldsymbol{x}_i) + q_i \log P(\text{benign}|\boldsymbol{x}_i) \\ &= -\frac{1}{N}\sum_{\boldsymbol{x}_i \in X} p_i \log y(\boldsymbol{x}_i) + q_i \log(1 - y(\boldsymbol{x}_i)), \end{aligned} \tag{4}$$

where X is the set of distinct feature vectors occurring in the test data, p_i and q_i are the numbers of adverse and benign outcomes for vector $\boldsymbol{x}_i$, and N is the total number of test set records. ANLL is minimized when the predicted risk of adverse outcome for profile $\boldsymbol{x}_i$ is

$$y(\boldsymbol{x}_i) = \frac{a_i}{a_i + b_i}. \tag{5}$$

3.3.1 Performance with 4 features

Although the 4 non-ICD code features `MumAge`, `Parity`, `CSections` and `NeoDxs` were encoded in a 12 dimensional binary vector, the 1-of-N representation of `MumAge` and `Parity` meant that only 126 distinct feature vectors appeared in the training data. Given that the actual number of patient records outnumbered that by three orders of magnitude, it is not surprising that the NetEns, LkpTab and LogReg models achieved close to maximum attainable discrimination in this setting.

By calculating MAD, we know that the performance of the different prediction models is limited by the 4-variable representation of the data, rather than shortcomings in the models themselves.

[4] Table 1 and Figure 3 show that maximum EAD is achieved with at least nine predictor variables.

Failure to progress												
Feature, ICD code and Range			**Order of addition**									
Parity		0-4	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
MumAge		0-6	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
CSections		0,1	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
SUPRV HIGH-RISK PREG NEC	V238	0,1	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
THREATENED LABOR NEC	6441	0,1	**6**	**6**	**5**	**5**	**5**	**5**	**5**	**5**	**6**	**8**
Height		0-2	**5**	**8**	**6**	**6**	**6**	**6**	**7**	**6**	**9**	**5**
ANEMIA IN PREGNANCY	6482	0,1	**7**	**5**	**10**	**9**	**7**	**9**	**6**	**10**	**5**	**7**
POOR FETAL GROWTH	6565	0,1	**9**	**11**	**8**	**12**	**10**	**12**	**8**	**7**	**8**	**6**
PREM SEPARATION PLACENTA	6412	0,1	13	**10**	18	**13**	**12**	**10**	12	**12**	12	23
HIGH HEAD AT TERM	6525	0,1	27	16	**12**	15	**13**	23	25	**9**	20	10
OTH PLACENTAL CONDITIONS	6567	0,1	22	14	**11**	28	**9**	**11**	22	29	**11**	18
ABNORMAL VULVA IN PREG	6548	0,1	19	15	19	22	**11**	20	**9**	32	21	14
NeoDxs		0,1	12	31	16	26	17	**13**	13	**8**	36	17
HEMORR FROM PLACENT PREV	6411	0,1	17	**9**	33	**8**	22	28	20	38	35	28
TRANSVERSE/OBLIQUE LIE	6523	0,1	28	17	30	18	26	24	19	**11**	14	11
ANTEPARTUM HEMORR NEC	6418	0,1	**10**	24	24	20	29	16	36	22	39	32
INFECTIV DIS IN PREG NEC	6478	0,1	31	27	28	16	15	25	17	25	17	20
FETAL DISPROPORTION NOS	6535	0,1	11	35	20	34	33	33	18	17	26	21
INDICAT CARE LAB/DEL NEC	6598	0,1	35	18	26	39	14	19	30	16	15	27
MALPOSITION NEC	6528	0,1	**8**	12	34	**7**	24	22	21	37	33	39
ABO ISOIMMUNIZATION	6562	0,1	32	22	13	19	34	26	29	18	22	15
MALPOSITION NOS	6529	0,1	14	13	17	27	18	38	23	20	37	35
OBESITY	2780	0,1	26	26	14	37	25	29	10	**13**	16	38
EXCESSIVE FETAL GROWTH	6566	0,1	24	38	22	21	20	42	11	15	25	13
PREGNANCY COMPL NOS	6469	0,1	37	29	25	31	21	27	27	24	23	31
CEPHALIC VERSION NOS	6521	0,1	29	34	23	17	35	**14**	16	27	18	16
CERVIX INCOMPET IN PREG	6545	0,1	18	20	**9**	14	19	21	38	21	27	22
PREG W POOR REPRODUCT HX	V235	0,1	30	19	35	24	28	30	15	23	30	26
UTERINE TUMOR IN PREG	6541	0,1	16	37	32	**11**	16	41	24	19	29	34
VOMITING COMPL PREG NEC	6438	0,1	25	21	31	32	31	15	14	14	19	33
POLYHYDRAMNIOS	6579	0,1	23	25	21	41	36	32	26	30	42	19
PREG W HX OF INFERTILITY	V230	0,1	20	30	37	23	37	**7**	34	33	40	29
DIABETES MELLIT IN PREG	6480	0,1	34	36	15	29	27	35	39	36	**10**	24
TRANS HYPERTENSION PREG	6423	0,1	40	32	39	36	**8**	34	35	42	13	12
RHESUS ISOIMMUNIZATION	6561	0,1	41	40	36	25	38	36	37	26	32	40
PREG W POOR OBSTETRIC HX	V234	0,1	36	23	29	35	39	17	28	34	24	37
MILD HYPEREMESIS GRAVID	6430	0,1	21	33	27	38	30	39	33	28	28	36
BONE DISORDER IN PREG	6487	0,1	33	39	41	40	23	40	32	31	41	25
THREATENED ABORTION	6400	0,1	38	42	40	43	42	37	42	40	31	41
PROLONGED PREGNANCY	6459	0,1	47	**7**	**7**	46	40	45	46	45	**7**	**9**
ABN GLUC TOLERAN IN PREG	6488	0,1	39	28	38	30	32	31	40	35	38	30
THREATEN PREMATURE LABOR	6440	0,1	15	41	43	42	44	18	41	39	43	42
MILD/NOS PRE-ECLAMPSIA	6424	0,1	42	45	42	44	46	43	44	43	45	45
EDEMA IN PREGNANCY	6461	0,1	43	43	44	33	41	44	31	41	34	44
HYPERTENS COMPL PREG NOS	6429	0,1	45	47	45	**10**	45	**8**	45	44	44	46
ThrptAbortions		0,1	44	44	46	47	47	46	43	46	46	43
Support		0,1	48	48	47	45	43	47	47	47	47	48
SpontAbortions		0,1	46	46	48	48	48	48	48	48	48	47
SocClass		0-6	49	49	49	49	49	49	49	49	49	49

Table 1: The discrete variables used to predict failure to progress, and the order in which they were added to the working set across 10 of the 20 separate runs of FEAD. Capitalized entries indicate ICD-9 code variables describing maternal conditions. Numbers in bold indicate the subset of features that attained maximum EAD in each run.

Average negative log-likelihood			
Model	**4 vars**	**35 vars**	**49 vars**
Test set limits	0.283	0.262	0.211
NetEns	0.284	0.280	0.280
NetMix	—	0.281	—
LogReg	0.285	0.282	0.281
LkpTab	0.284	0.284	0.296
Null model	0.328	0.328	0.328

Area under ROC curve			
Model	**4 vars**	**35 vars**	**49 vars**
Test set limits	0.764	0.808	0.883
NetEns	0.763	0.776	0.777
NetMix	—	0.773	—
LogReg	0.762	0.775	0.775
LkpTab	0.763	0.767	0.740
EAD	0.763	0.766	0.697
Null model	0.500	0.500	0.500

Table 2: The performance of each prediction model as a function of the number of predictor variables used by that model.

3.3.2 Performance with 35 features

As the representation of the SMR2 data becomes more complex, the ability to resolve individual test cases increases. This explains the non-decreasing MAD curve in Figure 3. However, our ability to make good predictions about the risk associated with each vector of patient characteristics does not necessarily increase. While the amount of training data remains the same as in the 4-feature model, the 35 variable representation increases the number of distinct feature vectors to 13 500. From this point onwards, the number of parameters in each prediction model (Figure 6) starts to have a noticeable impact on their performance.

When the ratio of training data to model parameters decreases beyond a certain point, the curse of dimensionality takes its toll on generalization performance. Figure 6 shows the exponential increase in parameters of both the LkpTab model and the histogram probability density model used to estimate EAD. As a consequence of this, the performance of EAD and the LkpTab model falls below that of the more biased LogReg and NetEns models when 35 features are used to represent the data. The number of unassigned bins (Figure 7) increases monotonically with the complexity of representation.

Each of the 20 networks in the NetMix model uses different subsets of 9–15 features to represent the data. In total, 35 features are used in the NetMix model. The performance of individual networks is shown in the boxed region of Figure 5. The prediction accuracy obtained by averaging the 20 outputs together is reported in Table 2 and is marginally less than that of the more complex NetEns model.

3.3.3 Performance with 49 features

When the full complement of features is used to represent the SMR2 data, EAD and the performance of the LkpTab model fall dramatically. The smoothing effect of the Dirichlet prior keeps the LkpTab model performance above that of the simple histogram density model. The number of parameters in the LogReg and NetEns models are, respectively, linear and quadratic functions of the number of features. As such, their generalization performance is not yet diminished by the high dimensional representation of the data.

Note that the performances of NetEns and LogReg are only slightly greater than the largest values of EAD observed in each run of the FEAD algorithm. The maximum EAD values ranged between 0.773 and 0.775 across the 20 runs, and occurred when the working set contained 9–15 variables.

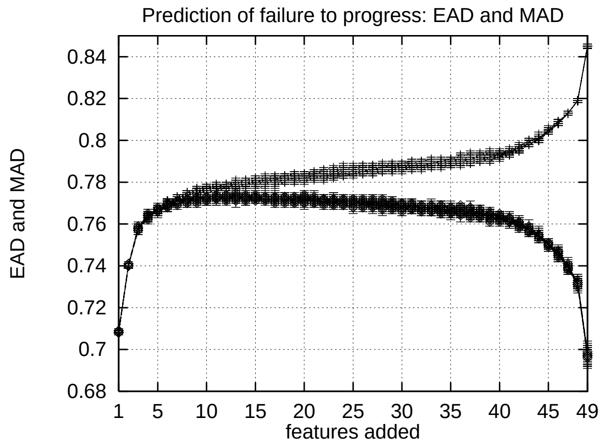

Figure 3: MAD (upper curves) and EAD (lower curves) for prediction of failure to progress, obtained over 20 runs of FEAD. Error bars show one standard deviation above and below each point estimate. Features are added, one at a time, from left to right, starting with a single feature model. The order of addition is shown in Table 1.

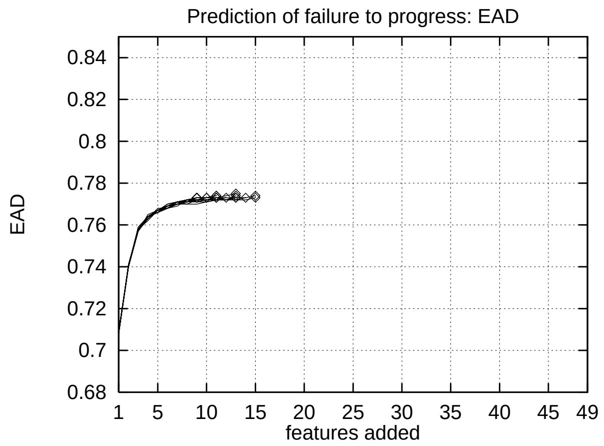

Figure 4: The EAD obtained over 20 runs of FEAD, up to the point of maximum EAD. The features corresponding to the points in this graph are shown in bold in Table 1

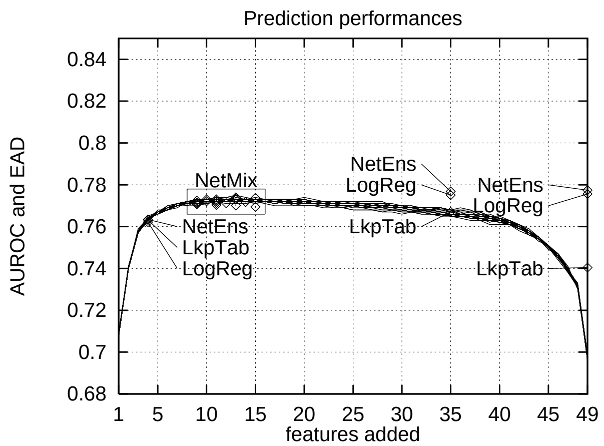

Figure 5: The test set performance of different models plotted against the number of features used by those models.

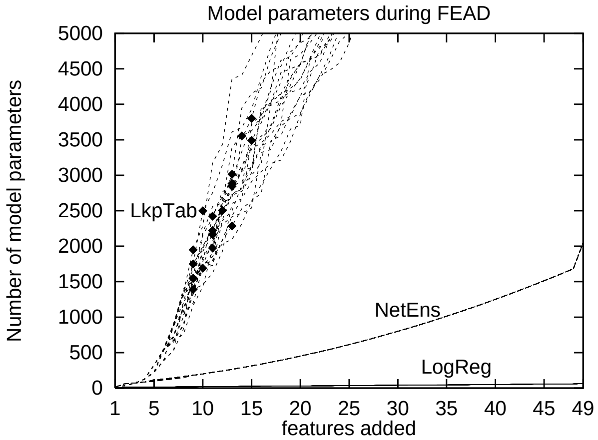

Figure 6: The number of parameters in the LkpTab, NetEns and LogReg models plotted against the features selected in twenty different runs of FEAD. The number of parameters in the LkpTab are equal to the number of bins in the histogram probability density model used to estimate EAD. The black points on the curves for the LkpTab model show the number of bins that were in the histogram probability density model when maximum EAD was obtained.

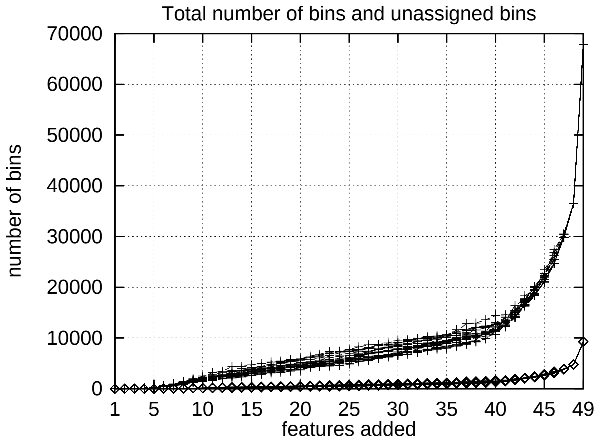

Figure 7: The number of bins, and the number of *unassigned* bins plotted across twenty different runs of FEAD. Recall that an unassigned bin is one that appears in the D^2 partition of the data but not in D^1 (see Section 2.3).

4 Discussion

The results presented in the previous section show that we can obtain good prediction performance with neural networks and logistic regression using features selected with the help of EAD. Furthermore, from a clinical perspective, the ranking of features shown in Table 1 accords with conventional obstetric wisdom. This is notable since the algorithm operated on database information alone, without recourse to expert opinion or prior assumptions.

In Section 4.1 we consider why it is extremely hard to go beyond this kind of empirical evidence for, or against, a given approach. Although we would like to provide a meaningful quantitative evaluation of EAD for feature selection, there are good reasons why this cannot be done without reference to a specific prediction model. It must be recognized that accurate prediction is as much a function of the representation of data as it is of the power of our prediction models, and that it would be misleading to consider the choice of representation in isolation.

Instead of fabricating toy problems where EAD leads to the selection of better features, faster than other methods, we present a qualitative assessment of the pros and cons of EAD as a feature selection criterion. By doing this, we hope to provide other researchers with a clear idea of where this approach could be put to good use.

4.1 Difficulties in assessing feature selection algorithms

Comparing feature selection algorithms is a tricky business. There are three reasons for this:

1. feature selection can be used as a tool to understand the problem at hand, and as a means to improve prediction performance. The varying importance of these different aims makes it hard to achieve a general comparison between algorithms;

2. if we perform feature selection to improve prediction performance, its effect is modulated by the prediction system that uses the selected features. Thus it is difficult to determine what degree of performance is attributable to the feature selection method, and what is attributable to the modelling technique;

3. in certain problem settings, other factors, such as execution speed or sensitivity to distributional assumptions, may be significant. This can prohibit a general comparison of algorithms.

There are two possible objectives in feature selection. One is to discover an underlying relationship between predictors and outcome so as to achieve a better *understanding* of the problem at hand. The other is to determine the variables that allow us to build the *most accurate* prediction model, given a finite amount of data. To some extent, these objectives are compatible, but it would be naïve to think that they are one and the same at the extremes. The 4-variable LogReg model of failure to progress has a straightforward interpretation but, in terms of log predictive error, it is clearly not the most accurate model. At the other extreme, the 49-variable NetEns model achieves marginally better prediction performance than all other models, but provides very little in the way of explanation. This highlights the difference between selecting features to *explain* a particular outcome, and selecting features to *predict* that outcome.

We cannot speak of feature selection to improve prediction performance without reference to some kind of prediction system. Since different prediction models may be more or less suited to a given task, it is hard to disentangle the effects of *feature selection* and *model bias* on prediction performance. These difficulties are compounded by the lack of a well accepted definition of bias and variance for 0/1 loss (*i.e.*, classification) problems [17].

Other factors may have a bearing on the utility of a feature selection algorithm. Some methods rely upon certain assumptions about the within-class distribution of data points, *e.g.*, methods that use Mahalanobis distance as a selection criterion [18]. In certain circumstances, these algorithms may be less attractive than more robust techniques. Some search algorithms (*e.g.*, branch-and-bound [19]) require monotonicity of the feature selection criterion function, *i.e.*, that the "goodness" of a subset of features can never be decreased by addition of new features. This assumption may also be violated in practice, which lessens the appeal of such techniques. Finally, a computationally intensive algorithm that selects excellent subsets of features may be impractical to use with large amounts of data and/or features. Each of these factors makes it difficult to decide which feature selection strategy is "best" in general.

So, there are a variety of reasons why head-to-head comparison of FEAD and other algorithms might not be fruitful. In the next sections, we consider the strengths and weaknesses of feature selection using EAD so the reader may obtain a better idea of situations where this approach would be applicable.

4.2 The strengths of EAD for feature selection

EAD has five main strengths as a feature selection criterion:

1. EAD provides a performance benchmark, rather than an abstract statistical measure of interclass separation. EAD estimates directly the performance we could expect to achieve with a histogram density model of the data. With statistical separation measures, it is necessary to build prediction models using the selected features before we can obtain an idea of the practical value of that data representation.

   The complexity of all other models will lie between that of the histogram density model and the null model (*i.e.*, the model that predicts constant risk). Thus, EAD is a sensible measure to evaluate in exploring the discriminative power of feature subsets.

   The idea of providing a simple, fast benchmark to which other, more complicated prediction strategies could be compared has also been put forward by Holte [20], though not in the context of feature selection.

2. EAD does not make distributional assumptions about the data. The histogram density model has sufficient flexibility to model any probability distribution on discrete valued input space. So, unlike logistic regression, for example, EAD has the ability to model *any* interactions that appear in the data without having to specify those interactions beforehand.

3. Because EAD is estimated on the basis of several random partitions of a dataset, different searches that use EAD as a criterion will return slightly different results. This is a natural consequence of the uncertainty about the features that are important in solving a given problem. This uncertainty arises because we are trying to estimate model parameters with a finite amount of data.

   As far as we are aware, most other approaches to feature selection do not perturb data that is used to estimate the "goodness" of feature subsets. By avoiding the issue of how sensitive a feature selection criterion is to different samples of the available data, these methods can present an over-confident statement of the discriminative power of variables. EAD straightforwardly accounts for uncertainty about the importance of different features.

4. EAD is well suited as a feature selection criterion for large amounts of sparse discrete valued data. Data sets with these characteristics are common in medical settings and in other data mining tasks (*e.g.*, credit card risk scoring).

5. EAD is an appropriate feature selection criterion in situations where our goal is *risk prediction* as opposed to *classification*. This is because EAD is based on the AUROC — a non-parametric measure of the discrimination achieved by a given input-output mapping — rather than classification rate.

   Unlike the Bayes error rate, EAD and MAD can be calculated without specifying a loss matrix. This means that these AUROC based discrimination measures may be especially useful in situations where the loss matrix is difficult to elicit or prone to variation.

   The AUROC has been rightfully criticised by Jiang *et al.* [21], who state that it "is not a meaningful summary of clinical diagnostic performance when high sensitivity must be maintained". We must stress that we use AUROC primarily to direct our search for discriminative subsets of features. Proper evaluation of models trained on the selected features requires additional measures of performance, including estimates of log predictive loss and probability calibration plots.

4.3 The weaknesses of EAD for feature selection

EAD was developed with a specific kind of dataset in mind; it is certainly not a suitable approach to feature selection in general. There are five significant shortcomings of EAD as a feature selection criterion:

1. EAD can give pessimistic estimates of the performance that could be achieved by biased models, especially when the ratio of data points to distinct feature vectors is low. However, this is an unavoidable risk of using one kind of model to predict the performance of another. The sensible way to deal with this issue is to monitor the performance obtained by alternate models using various feature subsets selected by EAD.

2. Using EAD as a feature selection criterion in conjunction with a heuristic search technique does not produce a single "best" subset of features. At the end of the search process, it is still very much up to the experimenter to decide which features should be carried through to the modelling stage.

   This criticism could be levelled at many feature selection methods, particularly those based on statistical measures of interclass separation. It could be argued that these algorithms are not designed to actually select features, but rather as a means to explore useful representations of a finite data set in a sensible manner.

3. Since EAD makes no distributional assumptions about the data, it can require massive amounts of data to work effectively. It is not a good idea to try to select features using EAD with small datasets — in those circumstances, prior knowledge (*i.e.*, a biased model) is necessary to achieve good performance. Unfortunately, choosing a suitably biased model is usually a difficult problem in itself.

4. EAD does not tell us very much about the underlying structure of a problem. We cannot gauge the significance of interactions between variables by when they leave, or enter, the subset maintained by the search heuristic. The problem of discovering an underlying model of high dimensional data (if one exists) is a fundamental challenge in data mining.

5. EAD has no practical extension to continuous valued features. While it may be feasible to quantize one or two continuous valued features (*e.g.*, maternal age, height), the curse of dimensionality is brought on exponentially fast in this way. Tree based methods like CHAID (see [22] for a review) get around this problem by recursively partitioning continuous data but such *axis-aligned splits* can perform poorly in the presence of correlated variables.

5 Summary

In this report, we have described EAD and shown how it can be used as a feature selection criterion. This approach was developed specifically as a means to find a discriminative representation of large amounts of sparse, discrete valued data. Unlike statistical measures of interclass separation, EAD provides a meaningful benchmark to which the performance of other prediction models can be compared. Furthermore, since EAD is estimated on the basis of random partitions of data, it gives us a straightforward way to characterise uncertainty about the discriminative power of different features.

One purpose of feature selection algorithms is to determine informative variables before embarking upon more computationally intensive modelling strategies. We have used a real-world problem to demonstrate the use of EAD in feature subset search: selection of discriminative predictors of failure to progress in labour. From a clinical point of view, the order in which features were selected agreed with their relevance to the obstetric condition being investigated.

The levels of EAD obtained in a forward selection search were compared to the performance of neural network, logistic regression and smoothed lookup table models of risk. As the ratio of distinct feature vectors to available data grew beyond a certain point, EAD gave increasingly pessimistic estimates of the performance achievable by more biased models (*i.e.*, logistic regression and neural networks).

We have explored, qualitatively, the advantages and disadvantages of using EAD as a feature selection criterion. The bias-free histogram density model upon which EAD is based has the power to capture complex interactions between variables given sufficient data. While it may take substantial amounts of data for such interactions to be determined, the representational power of the histogram density model makes EAD a suitable criterion to select features for prediction systems that can model those relationships.

Section 4 highlights the difficulties in quantitative comparison of feature selection algorithms. We have not attempted to resolve that issue in this report — the matter is substantial enough to warrant exploration in another paper. To the best of our knowledge, no systematic treatment of the factors involved in comparing feature selection algorithms has been published. We hope to address this topic thoroughly in a subsequent report.

A The LogReg model

This standard statistical method [16] was trained using all 540 905 cases in the learning set. LogReg has an important advantage over the other, more complicated models. The regression coefficients show the change in log odds that arises when a given risk factor is present. Positive coefficients indicate that the presence of a variable increases risk of adverse outcome, and *vice versa* for negative coefficients. While LogReg may not be the most accurate model of a complex nonlinear relationship, it can still be of significant explanatory value. The coefficients learned by the LogReg model are shown in Table 4.

B The NetEns and NetMix models

These models form predictions by averaging together the outputs of an ensemble of two layer feed-forward neural networks. Ensembles of predictors were used in an attempt to moderate over-confident risk estimates from individual predictors [23, 24]. In the absence of any theoretically optimal method to choose the ensemble size, we decided that 20 networks provided adequate modelling power with tolerable training time. For similar reasons, we chose the number of hidden units to be $\lfloor N/2 \rfloor$, where N is the number of inputs to a network.

To train individual networks within an ensemble, the learning set was randomly partitioned into a *training set* (containing two thirds of the 540905 cases) and a *validation set* (containing the remaining cases). A network was trained to minimize the *cross-entropy* error on the training set until its performance on the validation set was maximized, or the available CPU time was used. This *early-stopping* is described in [25] and was used instead of more sophisticated (and computationally intensive) second-order methods because of the large size of the SMR2 dataset.

The difference between the NetEns and NetMix models lies in the input information received by the networks in each ensemble. The 20 networks in the NetEns model received identical input information: all 35 variables listed in Table 3. Each network in the NetMix model received a different subset of those variables: the bullet points in each column of Table 3 show the variables present in each subset.

C The LkpTab model

The hierarchical Bayesian model, LkpTab, uses the entire learning set directly to make predictions. Let p_i, q_i be the number of adverse/benign cases in the learning set corresponding to patient profile $\boldsymbol{x}_i$. For these patient characteristics, the risk of adverse outcome predicted by LkpTab is

$$P(\text{adverse}|\boldsymbol{x}) = \frac{p_i + \alpha m}{p_i + q_i + \alpha}. \tag{6}$$

The two hyperparameters α and m smooth the estimates of $P(\text{adverse}|\boldsymbol{x}_i)$ appearing in each bin. When we have seen many cases corresponding to the profile $\boldsymbol{x}_i$, this estimate will be dominated by the likelihood $p_i/(p_i + q_i)$. If, however, few or no such cases have been seen, the estimate will be dominated by the Dirichlet prior and will be close, or equal to the hyper-parameter m. The hyper-parameter α corresponds to the number of $\boldsymbol{x}_i$ cases that must be present in the learning set before our estimate departs significantly from m. The approximate Bayesian methods used to determine α and m are described in [9] and yielded values of α between 7 and 13 and m between 0.083 and 0.094.

Variable	**Presence in NetMix model**																			
MumAge	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Parity	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Height	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
CSections	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
NeoDxs						•		•				•			•					
OBESITY								•							•					
THREATENED ABORTION											•							•		
HEMORR FROM PLACENT PREV		•		•								•								
PREM SEPARATION PLACENTA		•		•	•	•		•				•	•		•	•		•		•
ANTEPARTUM HEMORR NEC	•															•				
TRANS HYPERTENSION PREG					•															
MILD/NOS PRE-ECLAMPSIA																			•	
HYPERTENS COMPL PREG NOS				•		•														
THREATENED LABOR NEC	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
PROLONGED PREGNANCY		•	•						•	•	•				•					
PREGNANCY COMPL NOS																	•			
INFECTIV DIS IN PREG NEC												•								
DIABETES MELLIT IN PREG									•											
ANEMIA IN PREGNANCY	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
CEPHALIC VERSION NOS						•														
TRANSVERSE/OBLIQUE LIE								•					•							
HIGH HEAD AT TERM			•		•			•							•	•		•		•
MALPOSITION NEC	•			•								•								
UTERINE TUMOR IN PREG				•																
CERVIX INCOMPET IN PREG			•																	
ABNORMAL VULVA IN PREG					•		•											•		
RHESUS ISOIMMUNIZATION																		•		
ABO ISOIMMUNIZATION														•						
POOR FETAL GROWTH	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
EXCESSIVE FETAL GROWTH															•					
OTH PLACENTAL CONDITIONS			•		•	•			•						•			•		•
POLYHYDRAMNIOS													•							
PREG W HX OF INFERTILITY						•														
PREG W POOR REPRODUCT HX																		•		
SUPRV HIGH-RISK PREG NEC	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Table 3: Each of the 20 columns of bullet points indicates which variables were present in the input of a given network in the NetMix model.

Table 4: The coefficients learned by the LogReg model for 49-, 35- and 4-element feature sets. The model estimates risk for a given patient profile $\boldsymbol{x} = [x_1 x_2 \ldots x_K]$, $x_i \in \{0, 1\}$, as $y(\boldsymbol{x}) = 1/(1 + e^{-z})$, where $z = \beta_0 + \sum_{i=1}^{K} \beta_i x_i$. Polytomous variables (like `MumAge` and `Parity`) use a one-of-N encoding with the first (baseline) variable encoded by setting all relevent inputs to zero. The bars at the right of the table represent the values of the regression coefficients about 0.0, the centre line. A '$\star$' beside a regression coefficient indicates that, under the usual frequentist assumptions, 0.0 lies within the 95% confidence interval of that coefficient.

Variable	**Coeffs (49 vars)**	**Coeffs (35 vars)**	**Coeffs (4 vars)**
Constant	−1.714	−1.779	−2.019
MumAge <20	Baseline	Baseline	Baseline
MumAge 20-24	+0.263	+0.320	+0.321
MumAge 25-29	+0.579	+0.662	+0.654
MumAge 30-34	+0.819	+0.910	+0.888
MumAge 35-39	+1.077	+1.170	+1.131
MumAge >40	+1.370	+1.465	+1.417
MumAge Not Known	+0.338$\star$	+0.254$\star$	+0.266$\star$
Parity 0	Baseline	Baseline	Baseline
Parity 1	−2.122	−2.116	−2.097
Parity 2	−2.859	−2.861	−2.836
Parity 3	−3.092	−3.100	−3.063
Parity 4+	−2.980	−2.986	−2.929
SocClass I	Baseline		
SocClass II	−0.050		
SocClass IIIM	−0.045		
SocClass IIIN	−0.083		
SocClass IV	−0.059		
SocClass V	−0.025$\star$		
SocClass Not Known	−0.206		
Height < 155cm	Baseline	Baseline	
Height ≥ 155cm	−0.328	−0.322	
Height Not Known	−0.452	−0.453	
Support	+0.081		
SpontAbortions	+0.074		
ThrptAbortions	−0.073		
CSections	+1.969	+1.966	+1.939
NeoDxs	−0.610	−0.630	−0.652
OBESITY	+0.640	+0.669	
THREATENED ABORTION	−0.308	−0.318	
HEMORR FROM PLACENT PREV	−0.527	−0.525	
PREM SEPARATION PLACENTA	−1.249	−1.252	
ANTEPARTUM HEMORR NEC	+0.605	+0.598	
TRANS HYPERTENSION PREG	+0.253	+0.252	
MILD/NOS PRE-ECLAMPSIA	+0.130	+0.136	
HYPERTENS COMPL PREG NOS	+0.210	+0.219	
MILD HYPEREMESIS GRAVID	−0.242		
VOMITING COMPL PREG NEC	+0.796		
THREATEN PREMATURE LABOR	−0.151		
THREATENED LABOR NEC	−0.877	−0.884	
PROLONGED PREGNANCY	+0.179	+0.185	
EDEMA IN PREGNANCY	+0.104		
PREGNANCY COMPL NOS	+0.217	+0.211	
INFECTIV DIS IN PREG NEC	−0.545	−0.560	
DIABETES MELLIT IN PREG	+0.470	+0.484	

...continued from previous page.

Variable	Coeffs (49 vars)	Coeffs (35 vars)	Coeffs (4 vars)
ANEMIA IN PREGNANCY	+0.429	+0.424	
BONE DISORDER IN PREG	−0.248		
ABN GLUC TOLERAN IN PREG	+0.259		
CEPHALIC VERSION NOS	+0.927	+0.935	
TRANSVERSE/OBLIQUE LIE	+0.532	+0.532	
HIGH HEAD AT TERM	+0.925	+0.931	
MALPOSITION NEC	+0.935	+0.937	
MALPOSITION NOS	+0.397		
FETAL DISPROPORTION NOS	+1.581		
UTERINE TUMOR IN PREG	+0.230⋆	+0.225⋆	
CERVIX INCOMPET IN PREG	−0.157⋆	−0.164⋆	
ABNORMAL VULVA IN PREG	+0.425	+0.439	
RHESUS ISOIMMUNIZATION	−0.211	−0.213	
ABO ISOIMMUNIZATION	−0.375⋆	−0.384⋆	
POOR FETAL GROWTH	−0.495	−0.504	
EXCESSIVE FETAL GROWTH	+0.880	+0.882	
OTH PLACENTAL CONDITIONS	−1.161	−1.175	
POLYHYDRAMNIOS	+0.465	+0.473	
INDICAT CARE LAB/DEL NEC	−0.003⋆		
PREG W HX OF INFERTILITY	+0.302	+0.312	
PREG W POOR OBSTETRIC HX	−0.400		
PREG W POOR REPRODUCT HX	−0.121⋆	−0.113⋆	
SUPRV HIGH-RISK PREG NEC	−4.259	−4.251	

References

[1] World Health Organization, *International Classification of Diseases, Ninth Revision.* Geneva: World Health Organization, 1975.

[2] W. Siedlecki and J. Sklansky, "On automatic feature selection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 2, pp. 197–220, 1988.

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

[4] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, pp. 387–415, 1975.

[5] D. Green and J. Swets, *Signal Detection Theory and Psychophysics.* New York: John Wiley, 1966.

[6] C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[7] D. R. Lovell, C. R. Dance, M. Niranjan, R. W. Prager, and K. J. Dalton, "Limits on the discrimination possible with discrete valued data, with application to medical risk prediction," Tech. Rep. CUED/F-INFENG/TR243, Cambridge University Engineering Department, January 1996.

[8] J. Hanley and B. McNeil, "Maximum attainable discrimination and the utilization of radiologic examinations," *Journal of Chronic Disease*, vol. 35, pp. 601–611, 1982.

[9] D. J. C. MacKay and L. Peto, "A hierarchical Dirichlet language model," *Natural Language Engineering*, vol. 1, no. 3, pp. 1–19, 1995.

[10] S. M. Weiss and C. A. Kulikowski, *Computer systems that learn: classification and prediction methods from statistics, neural networks, machine learning, and expert systems.* San Mateo: Morgan Kaufmann, 1991.

[11] J. Kittler, "Features set search algorithms," in *Pattern Recognition and Signal Processing* (C. H. Chen, ed.), pp. 41–60, The Netherlands: Sijthoff and Noordhoff, 1978.

[12] D. R. Lovell, C. R. Dance, M. Niranjan, R. W. Prager, and K. J. Dalton, "Using upper bounds on discrimination to select discrete valued features," in *Neural Networks for Signal Processing VI* (S. Usui, Y. Tohkura, S. Katagiri, and E. Wilson, eds.), (Keihanna, Japan), pp. 233–242, 1996.

[13] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[14] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, pp. 1–58, 1992.

[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: the Art of Scientific Computing*. Cambridge University Press, 2nd ed., 1992.

[16] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York; Chichester: Wiley, 1989.

[17] D. H. Wolpert, "On bias plus variance," *Neural Computation*, vol. 9, no. 6, 1997.

[18] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[19] P. M. Narandra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. C-26, no. 9, pp. 917–922, 1977.

[20] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63–91, 1993.

[21] Y. L. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology*, vol. 201, no. 3, pp. 745–750, 1996.

[22] D. M. Hawkins and G. V. Kass, "Automatic interaction detection," in *Topics in Applied Multivariate Analysis* (D. M. Hawkins, ed.), pp. 267–300, Cambridge University Press, 1982.

[23] D. MacKay, *Bayesian methods for adaptive modelling*. PhD thesis, California Institute of Technology, 1991.

[24] M. Perrone, "Averaging techniques for neural networks," in *The Handbook of Brain Theory and Neural Networks*, MIT Press, 1993.

[25] C. E. Rasmussen, *Evaluation of Gaussian processes and other methods for non-linear regression*. PhD thesis, University of Toronto, 1996.