Design, construction and evaluation of systems to predict risk in obstetrics*

D. R. Lovell, B. Rosario, M. Niranjan, R. W. Prager,
K. J. Dalton†, R. Derom‡ and J. Chalmers*

Cambridge University Engineering Department,
Trumpington St., Cambridge CB2 1PZ, UK

†Cambridge University Department of Obstetrics and Gynaecology,
Rosie Maternity Hospital, Cambridge, CB2 2SW, UK

‡Katholieke Universiteit Leuven, Center for Human Genetics,
Herestraat 49, B-3000 Leuven, Belgium

*Information and Statistics Division of the National Health
Service in Scotland, Trinity Park House, Edinburgh EH5 3SE, UK.

Abstract

We present a systematic, practical approach to developing risk prediction systems, suitable for use with large databases of medical information. An important part of this approach is a novel feature selection algorithm which uses the area under the receiver operating characteristic (ROC) curve to measure the expected discriminative power of different sets of predictor variables. We describe this algorithm and use it to select variables to predict risk of a specific adverse pregnancy outcome: failure to progress in labour. Neural network, logistic regression and hierarchical Bayesian risk prediction models are constructed, all of which achieve close to the limit of performance attainable on this prediction task. We show that better prediction performance requires more discriminative clinical information rather than improved modelling techniques. It is also shown that better diagnostic criteria in clinical records would greatly assist the development of systems to predict risk in pregnancy.

Keywords: Receiver operating characteristic (ROC); Feature selection; Risk prediction in pregnancy; Failure to progress; Neural networks.

1 Introduction

Hospital maternity databases are a rich source of information about the health status of mothers and babies. We investigate the potential of this data to be used to predict the risk of a bad outcome for mother and child, specifically, failure to progress in labour.

In practical terms, an accurate obstetric risk prediction system could give clinicians an early indication of danger, thereby allowing enough time for medical intervention or closer monitoring of the pregnancy. While the medical aspect of this research is important, the central aim of this paper is to present a systematic, practical approach to developing risk prediction systems.

*Accepted for publication in *International Journal of Medical Informatics*.

In Section 2 we describe a novel method to help select items of information relevant to the prediction of risk of a given outcome. By calculating the maximum extent to which any system can discriminate between cases with good and bad outcome, the proposed technique provides a useful benchmark for prediction system performance. Section 3 shows how this method was used in designing systems to predict failure to progress in labour for a substantial obstetric database. The construction and evaluation of five different risk prediction models is described in the remainder of Section 3. The paper concludes with a discussion of the performance achieved by the individual models and looks at whether those systems could be used in practice.

2 Selection of discriminative predictors

In practice, it is difficult to decide which pieces of information should be used in making predictions of risk [1]. There are, however, compelling practical and theoretical reasons why such decisions about *feature selection* should be made.

The performance of the Bayes classifier can never be improved by *omission* of information [2] but, in practice, we are unable to build this classifier because we do not know the actual class conditional probabilities that characterise our data. Instead, we build a parameterised model of the system we wish to make predictions about, and use training data to obtain estimates of the model's parameters. With a finite amount of data, there is a trade off between the model's complexity and the accuracy with which we can estimate its parameters. Omission of one or more predictors simplifies our model and can lead to more accurate parameter estimates, thereby improving the performance of our non-ideal classifier [3, 4, Section 3.8].

From a practical perspective, the number of available predictors is governed by the number of observations we can actually make. In a medical setting, there are so many possible tests and measurements that we must determine the most informative of these to maximise the wellbeing of the patient, and keep demands on health care resources within reason.

2.1 Problem setting

We wish to make an accurate forecast of the probability of some *adverse* patient outcome, given certain patient characteristics. We refer to a case predicted to have adverse outcome as *positive*, and a case predicted to have *benign* outcome as *negative*.

We consider only discrete valued predictors of outcome since this allows us to estimate the risk associated with each combination of patient characteristics. Each distinct combination of predictor values is referred to as a "bin" or "patient profile". While the number of *possible* bins grows combinatorially with the number of predictors, this representation of the data only stores outcomes of profiles that *actually appear* in the data.

In ideal circumstances, each patient profile would be associated with a single class of outcome, allowing us to discriminate perfectly between adverse and benign cases. In reality, cases with identical predictors often go on to have different outcomes. Here, the *receiver operating characteristic* (ROC) curve [5] can provide a non-parametric measure of the degree of discrimination achieved by a risk prediction system (see Figure 1).

This paper deals with risk prediction systems that associate a continuous output (a *risk estimate*) with each patient profile. Arrangement of a database

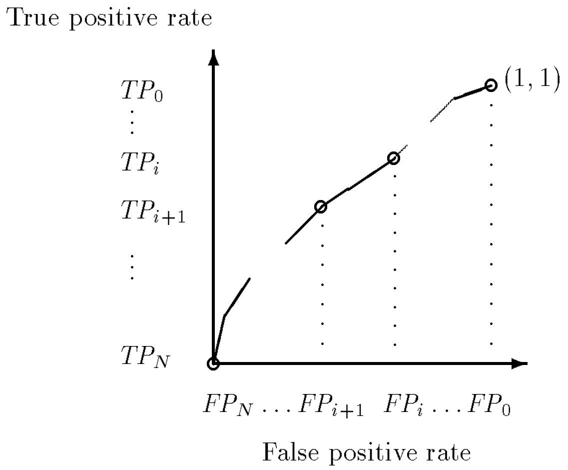

Figure 1: These sections of an ROC curve show true and false positive rates across the thresholds, $t_0, \ldots, t_N$, separating bins, $B_1, \ldots, B_N$, in an ordered list. Straightforward methods to calculate the area under this curve (and its standard error) are presented in [6] and [7].

of patient profiles in ascending order of estimated risk can be represented as

$$\begin{array}{ccccccccc}
 & B_1 & & B_2 & & & & B_N & \\
 & \boxed{\begin{matrix} a_1 \\ b_1 \end{matrix}} & & \boxed{\begin{matrix} a_2 \\ b_2 \end{matrix}} & & \ldots & & \boxed{\begin{matrix} a_N \\ b_N \end{matrix}} & \\
\uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\
t_0 & & t_1 & & t_2 & & t_{N-1} & & t_N
\end{array}$$

where bin B_i contains a_i adverse outcomes and b_i benign outcomes, and thresholds $t_0, \ldots, t_N$ separate each bin in the list. For any two bins in this order, the bin associated with the higher estimated risk appears to the right of the other bin. Using this representation, we define

$$\text{true positive rate at } t_i\text{:} \quad TP_i = \frac{\text{number of adverse cases right of } t_i}{\text{total number of adverse cases}}$$

$$\text{false positive rate at } t_i\text{:} \quad FP_i = \frac{\text{number of benign cases right of } t_i}{\text{total number of benign cases}}.$$

The ROC curve plots TP_i versus FP_i across all thresholds t_i. Furthermore, the *area under the ROC curve* measures how well a prediction system discriminates between two classes of outcome [5].

Unlike most probabilistic measures of inter-class separation, the area under the ROC curve has a number of physical interpretations [5, Chapter 2]. We adopt the following interpretation: given a randomly chosen case with adverse outcome, and a randomly chosen case with benign outcome, the probability that a risk prediction system correctly decides which of these cases was at higher risk is equal to the area under the system's ROC curve. Throughout the remainder of this paper, we shall use the terms "discrimination" and "area under the ROC curve" synonymously.

Note that by using discrete valued features to represent data, we could also employ Bayes risk as a measure of discrimination instead of the area under the ROC curve [8, Section 3.1]. However, for problems in which the cost of misclassification can vary, or is difficult to specify (as is often the case in medical risk prediction), it is unclear how the Bayes risk could be evaluated.

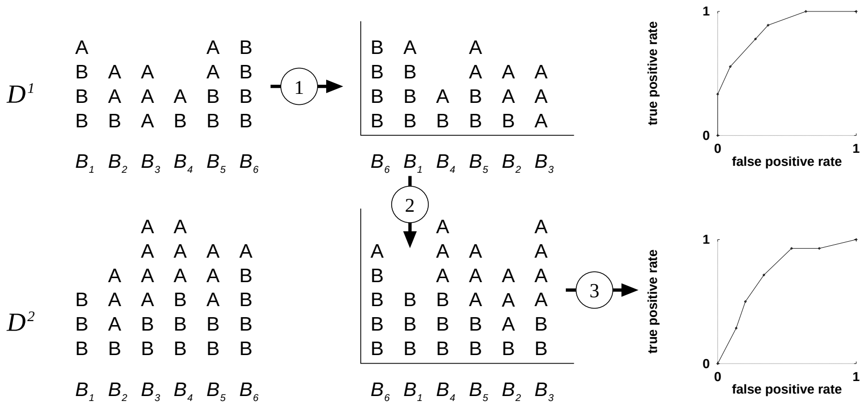

Figure 2: To estimate the discrimination we can *expect* to obtain with a given representation of a specific dataset, we randomly partition our dataset into two disjoint subsets, D^1 and D^2. The D^1 bins, B_1–B_6, are placed in MAD order (1). The D^1 MAD ordering is applied to the bins in the D^2 subset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

2.2 Maximum attainable discrimination

It is possible to calculate the maximum extent to which one can discriminate between adverse and benign outcomes in a given set of bins. From the definitions of true and false positive rates we have that

$$
\begin{aligned}
TP_i &= TP_{i-1} - \frac{a_i}{\text{total number of adverse cases}} \\
FP_i &= FP_{i-1} - \frac{b_i}{\text{total number of benign cases}}.
\end{aligned}
$$

Since the total numbers of adverse and benign cases 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 a_i/b_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, the line segments that make up the ROC curve appear, from left to right, in decreasing order of slope, and the area under the curve is maximized (see step (1) in Figure 2) [5, 9, 10].

This *maximum attainable discrimination* — or *MAD*, as Hanley and McNeil refer to it [10] — 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 4, the MAD helps determine whether a prediction system's performance is limited by an inherent lack of separability in the data, or shortcomings of the prediction model.

2.3 Expected attainable discrimination

We wish to establish how well a particular set of patient information allows us to discriminate between adverse and benign cases in *new* data. A classic way to estimate the *expected* performance of a system is to reserve, or "hold out" data for testing [4]. Given two sets of data, D^1 and D^2, drawn from the population we wish to classify, the *expected attainable discrimination* (EAD) can be estimated as follows (see Figure 2):

1. Obtain the MAD with D^1 by placing its bins in ascending order of likelihood ratio 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 the second data set but not the first are inserted randomly into the L^2 ordering. This is not the only way to deal with these "unassigned" bins (the Bayesian approach described in [11] could also be used), however this method was used throughout the experiments described in this paper.

In practice, the estimates of expected discrimination are formed by randomly partitioning a data set, D, into two disjoint subsets, D^1, D^2. Estimates from a series of such random partitions are averaged to form a *rotation error* estimate [3, p. 216] which mitigates the effects of sampling variability.

2.4 Searching for discriminative subsets of predictors

An exhaustive search of all possible subsets of predictors is the only way to guarantee discovery of the representation which maximizes EAD[1] [3]. There are, however, a number of heuristic search methods (*e.g.*, backwards, forwards and stepwise selection [3]) that can help us explore the subsets of predictors with large EAD.

In Section 3 a *backwards elimination* algorithm is used to remove predictors in increasing order of their discriminative power. (For convenience, we refer to Backwards elimination using EAD as BEAD.)

Commencing with the *saturated* model (*i.e.*, using all N available predictors to represent the data), the EADs of all possible models with $(N-1)$ predictors are evaluated. The $(N-1)$ variable model with the highest EAD is selected and all its possible $(N-2)$ predictor sub-models are evaluated. This process continues until we have a model that contains a single predictor variable. Because there is a degree of random variation in our EAD estimates, BEAD should be performed a number of times to reveal *trends* in the order in which predictors are removed from the model.

What are the advantages of BEAD in comparison to backwards elimination using, say, logistic regression? Foremost, is that by using EAD we automatically model *all* variable interactions appearing in the data. With logistic regression, we must explicitly state those variable interactions which we suspect may be important — a difficult task when there are large numbers of variables with uncertain degrees of interaction.

Computationally, the sorting process required to estimate EAD is more efficient than the optimization step needed to fit an equivalent logistic regression model (although this advantage becomes less significant when several estimates of EAD are averaged together). Accurate evaluation of EAD requires sufficient data for us to make good estimates of the risk associated with each combination of patient characteristics. Thus, the techniques described here are particularly well suited to large medical databases in which the number of different patient profiles that actually appear is small enough for there to be, say, more than 10 examples in most bins.

It has recently come to our attention that the feature selection algorithm we propose has much in common with the *automatic interaction detection* (AID) methods described by Morgan and Sonquist [12] (see [13] for a review). The main difference between the approach we have presented and the AID family of

[1] It should be remembered that EAD is still estimated from a finite sample, hence the subset of predictors that maximizes EAD is not necessarily the best for the population from which that sample was drawn.

algorithms is in our use of hold-out data to assess the discriminative power of subsets of features; methods like CHAID [14] rely upon in-sample tests of significance of association to determine potentially useful predictors. Although the possibility of finding spurious interactions with CHAID has been investigated empirically and appears to be small, estimation of EAD could further improve rejection of such artifacts, while providing a practical prediction performance benchmark.

3 Prediction of failure to progress in labour

In this section, we show how we have used MAD and EAD in designing systems to predict risk of adverse pregnancy outcome (APO). These methods were developed under the Quality Assurance in Maternity Care (QAMC) project, (1) to place upper bounds on the accuracy of prediction of different APOs and (2) to discover which features best determine risk of these APOs. One of the outcomes considered in the QAMC project is *Cæsarean section for failure to progress in labour* and in this section we discuss an investigation into the precursor to that outcome: failure to progress in labour. Put simply, a failure to progress means that the natural course of labour has stalled before delivery, placing the well-being of both mother and child at risk. In short, our aim was to determine how accurately a prediction of failure to progress could be made on the basis of information available prior to labour.

3.1 Scottish Morbidity Record data

We used data from the Scottish Morbidity Record (known as the SMR2) on 771571 singleton births that occurred between 1980–91. Cases in which elective Cæsarean section or breech presentation occurred were removed from the dataset[2]. 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.

3.2 Selection of predictor variables

If we wish to build a prediction system for use at a certain stage of a pregnancy, we can use only the maternal information available at that stage to train that system. The SMR2 data does not give precise details of when a diagnosis was made, so we consider information available at the onset of labour since it is fairly simple to establish which diagnoses were made prior to that.

Each SMR2 record can use up to 6 *ICD-9* codes [15] to describe maternal condition, and the learning set contains instances of over 300 different maternal conditions occurring before onset of labour. As mentioned in Subsection 2.4, accurate estimates of EAD require the majority of bins to hold more than 10 cases, so, before applying BEAD, we used univariate statistics to reduce the potential predictors to a manageable number. Using Cramer's V statistic [16], the predictors were ranked in order of their strength of association with failure to progress, and an arbitrary cutoff was used to select the 49 conditions most strongly associated with the outcome. Obviously, this univariate screening process is not the optimal approach to choosing promising variables for multivariate prediction, however, we are not aware of a more effective scheme that could be practically applied to such a large set of candidate variables.

[2] 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.

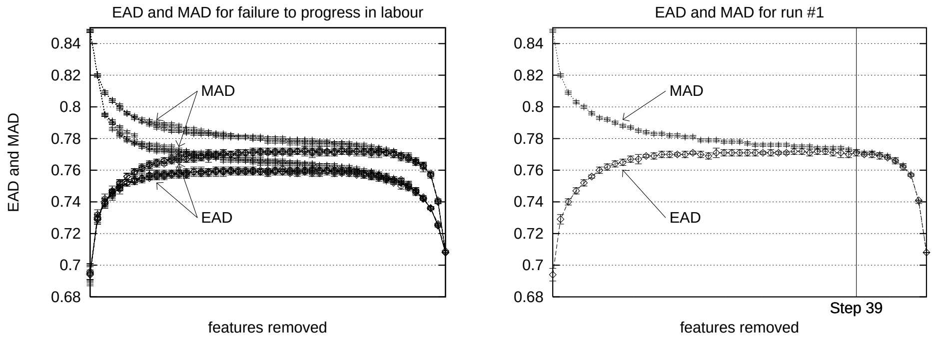

Figure 3: The left graph shows the MAD (upper curves) and EAD (lower curves) for prediction of failure to progress, obtained over 10 runs of BEAD. For clarity, the upper and lower curves of the right graph show the MAD and EAD obtained on the first run of the algorithm. Features are removed, one at a time, from left to right, starting with the saturated model. The order of removal is shown in Table 1. The error bars show one standard deviation above and below each point estimate.

The vertical line at the 39th step of the right graph shows the point at which further removal of features resulted in EAD below the peak value of 0.772. All features to the right of this point were used in building actual risk prediction models.

Next, we used BEAD to rank the 49 predictors in order of their expected ability to discriminate between instances of failure to progress and normal labour. As explained in Subsection 2.3, EAD is estimated on the basis of several random partitions of the learning data. Hence, each run of BEAD will tend to produce slightly different results even though the entire learning data set remains fixed. Variation in results implies a degree of uncertainty about the true relation between predictors and outcome which, in turn, arises because of the finite amount of data used to estimate this relationship. For this reason, it is important to look for *trends* in the order in which features are removed to identify features essential to achieve good discrimination.

The results of 10 separate runs of BEAD are shown in Figure 3 and Table 1. There are clearly two sets of curves in the graph of Figure 3. The curves with lower MAD and EAD represent 6 runs of BEAD in which the `MumAge` feature was removed at step 2 of the elimination process (see Table 1). (The early stages of the algorithm are when the estimates of EAD are least accurate, due to the relatively small number of cases in many of the bins.) In comparison to the 4 runs in which `MumAge` was retained until step 48, the lower levels of EAD and MAD indicate that `MumAge` is actually a discriminative feature which should be retained[3]. Taking this into account, we selected all features whose removal caused a strict *decrease* in EAD in the remainder of any of those 4 runs of BEAD (see the right graph of Figure 3). This criterion resulted in the selection of 24 features shown in Table 2. For the training of the prediction systems, these features were encoded in a 32 element binary feature vector.

3.3 Construction of prediction systems

The experiments described in this section aimed to "systematically entertain a range of possible models" [17, p. 378] with a view to establishing which model(s) were best suited to the prediction of failure to progress in labour. Five prediction

[3] This has been confirmed by additional experiments using *forward* selection of variables.

Failure to progress in labour										
Condition	**Order of removal**									
SocClass	1	1	1	1	1	1	1	1	1	1
Height	2	7	3	5	2	4	4	2	2	4
SpontAbortions	4	4	5	3	3	3	6	3	3	3
ThrptAbortions	5	6	7	4	5	6	5	4	9	6
PROLONGED PREGNANCY	6	3	6	8	7	12	3	6	4	5
MILD/NOS PRE-ECLAMPSIA	13	5	4	7	8	5	7	7	10	28
EDEMA IN PREGNANCY	10	8	9	11	10	8	12	9	8	13
THREATENED ABORTION	20	14	14	6	11	26	31	15	7	9
PREG W HX OF INFERTILITY	9	12	8	26	19	25	13	16	20	10
TRANS HYPERTENSION PREG	24	17	26	12	13	16	15	10	26	14
PREGNANCY COMPL NEC	12	29	15	24	15	20	33	8	13	11
BONE DISORDER IN PREG	38	9	17	21	14	13	16	25	22	12
THREATEN PREMATURE LABOR	16	40	18	13	6	14	35	18	11	22
MumAge	48	2	2	2	48	2	2	48	48	2
TRANSVERSE/OBLIQUE LIE	33	33	13	10	25	9	11	22	27	27
PREG W POOR OBSTETRIC HX	26	25	10	16	35	30	8	19	21	21
RHESUS ISOIMMUNIZATION	31	13	19	17	16	21	18	35	28	16
CEPHALIC VERSION NOS	19	10	16	34	23	28	17	23	12	34
ABN GLUC TOLERAN IN PREG	17	16	25	32	22	29	10	27	25	17
POLYHYDRAMNIOS	21	11	29	27	34	18	28	13	38	7
HYPERTENS COMPL PREG NOS	7	19	43	42	9	41	9	11	6	43
MALPOSITION NOS	34	23	34	14	39	15	26	14	19	15
ANTEPARTUM HEMORR NEC	14	22	24	29	12	23	29	31	30	20
ABO ISOIMMUNIZATION	22	35	35	31	31	7	24	20	33	8
FETAL DISPROPORTION NOS	8	20	27	25	36	32	34	28	14	24
HEMORR FROM PLACENT PREV	25	27	21	30	24	27	20	21	34	19
DIABETES MELLIT IN PREG	32	28	31	20	33	33	14	17	15	26
INDICAT CARE LAB/DEL NEC	15	26	32	15	30	19	25	38	17	33
EXCESSIVE FETAL GROWTH	30	37	12	22	17	10	36	30	18	39
HX-PENICILLIN ALLERGY	23	15	20	23	32	24	22	29	41	23
ABNORMAL VULVA IN PREG	27	32	28	18	18	31	19	26	29	25
VOMITING COMPL PREG NEC	29	24	23	33	28	11	30	24	31	32
UTERINE TUMOR IN PREG	28	21	30	9	20	22	32	34	35	35
GRAND MULTIPARITY	37	18	22	28	21	17	27	32	36	31
OTH PLACENTAL CONDITIONS	18	31	11	19	27	34	39	36	32	38
Support	3	47	47	47	4	47	47	5	5	47
ELDERLY PRIMIGRAVIDA	11	41	40	38	26	40	40	37	16	18
INTRAUTERINE DEATH	41	34	33	35	40	38	37	12	23	30
PREM SEPARATION PLACENTA	39	42	36	40	29	36	38	33	40	29
HIGH HEAD AT TERM	36	30	38	36	41	37	23	43	42	37
NeoDxs	42	43	39	41	37	39	21	40	24	40
MALPOSITION NEC	35	36	37	37	38	35	41	41	37	36
OTH CURRENT COND OF PREG	40	39	42	39	42	42	43	39	39	42
POOR FETAL GROWTH	43	38	41	43	45	43	42	42	43	41
ANEMIA IN PREGNANCY	44	44	45	45	44	45	44	44	44	45
THREATENED LABOR NEC	45	45	44	44	43	44	45	45	45	44
SUPRV HIGH-RISK PREG NEC	46	46	46	46	46	46	46	46	46	46
CSections	47	48	48	48	47	48	48	47	47	48
Parity	49	49	49	49	49	49	49	49	49	49

Table 1: The order in which features were removed from the saturated model in ten different runs of the backwards elimination algorithm. Variables listed entirely in upper case are abbreviations for ICD-9 codes in the maternity record.

Selected predictors of failure to progress			
ICD-9	Predictor	Incidence	
	CSections	21806	4.031%
	MumAge		
	NeoDxs	11868	2.194%
	Parity		
6412	PREM SEPARATION PLACENTA	3357	0.621%
6418	ANTEPARTUM HEMORR NEC	1321	0.244%
6441	THREATENED LABOR NEC	22421	4.145%
6482	ANEMIA IN PREGNANCY	51727	9.563%
6489	OTH CURRENT COND OF PREG	14573	2.694%
6525	HIGH HEAD AT TERM	2728	0.504%
6528	MALPOSITION NEC	4115	0.761%
6529	MALPOSITION NOS	309	0.057%
6541	UTERINE TUMOR IN PREG	402	0.074%
6561	RHESUS ISOIMMUNIZATION	8752	1.618%
6564	INTRAUTERINE DEATH	1914	0.354%
6565	POOR FETAL GROWTH	25243	4.667%
6566	EXCESSIVE FETAL GROWTH	2694	0.498%
6567	OTH PLACENTAL CONDITIONS	3510	0.649%
6579	POLYHYDRAMNIOS	2549	0.471%
6594	GRAND MULTIPARITY	640	0.118%
6595	ELDERLY PRIMIGRAVIDA	1800	0.333%
6598	INDICAT CARE LAB/DEL NEC	5049	0.933%
V140	HX-PENICILLIN ALLERGY	197	0.036%
V238	SUPRV HIGH-RISK PREG NEC	9351	1.729%

Table 2: A list of the predictors selected on the basis of 10 runs of the backwards elimination algorithm. Where appropriate, ICD-9 codes are shown in the leftmost column (full definition of these codes can be found in [15]), and the incidence of maternal conditions in the learning set of 540905 cases is shown in the two rightmost columns. The maternal age variable `MumAge` is discretised into 7 groups: < 20, 20-24, 25-29, 30-34, 35-39, > 40, and "not known". The variable `Parity` can take on values 0, 1, 2, 3, ≥ 4.

models were investigated:

Logistic regression (LogReg) This standard statistical method [18] was trained using all 540905 cases in the learning set.

Ensemble of logistic regressors (LogEns) Each member of this ensemble of 50 logistic regressors was trained on a unique random selection of two thirds of the records from the learning set.

Ensemble of neural networks (NetEns) Each two layer feed-forward network in this ensemble of 30 networks was trained on a random selection of two thirds of the learning set records.

Ensemble of neural networks with direct weights (DirEns) There were 44 of these networks (whose structure we describe below), each trained using a random selection of two thirds of the learning set records.

Lookup table with Dirichlet prior (LkpTab) This hierarchical Bayesian model is described fully in [11].

LogReg has an important advantage over the other, more complicated models. The regression coefficients show the change in log odds (Equation (3) below)

Coeff.	Variable	Value
β_0		-1.987454
	MumAge < 20	Baseline
β_1	MumAge 20-24	$+0.315755$
β_2	MumAge 25-29	$+0.648367$
β_3	MumAge 30-34	$+0.887493$
β_4	MumAge 35-39	$+1.124512$
β_5	MumAge > 40	$+1.415986$
β_6	MumAge Not Known	$+0.250468$
	Parity 0	Baseline
β_7	Parity 1	-2.106965
β_8	Parity 2	-2.844174
β_9	Parity 3	-3.075124
β_{10}	Parity 4+	-2.948162
β_{11}	CSections	$+1.950169$
β_{12}	NeoDxs	-0.606717
β_{13}	PREM SEPARATION PLACENTA	-1.225588
β_{14}	ANTEPARTUM HEMORR NEC	$+0.596585$
β_{15}	THREATENED LABOR NEC	-0.927686
β_{16}	ANEMIA IN PREGNANCY	$+0.432109$
β_{17}	OTH CURRENT COND OF PREG	-0.776385
β_{18}	HIGH HEAD AT TERM	$+0.968416$
β_{19}	MALPOSITION NEC	$+0.929143$
β_{20}	MALPOSITION NOS	$+0.425813$
β_{21}	UTERINE TUMOR IN PREG	$+0.202529$
β_{22}	RHESUS ISOIMMUNIZATION	-0.250797
β_{23}	INTRAUTERINE DEATH	-2.032393
β_{24}	POOR FETAL GROWTH	-0.494424
β_{25}	EXCESSIVE FETAL GROWTH	$+0.899809$
β_{26}	OTH PLACENTAL CONDITIONS	-1.230844
β_{27}	POLYHYDRAMNIOS	$+0.531160$
β_{28}	GRAND MULTIPARITY	$+0.358108$
β_{29}	ELDERLY PRIMIGRAVIDA	$+0.325693$
β_{30}	INDICAT CARE LAB/DEL NEC	$+0.008482$
β_{31}	HX-PENICILLIN ALLERGY	$+0.693684$
β_{32}	SUPRV HIGH-RISK PREG NEC	-4.370316

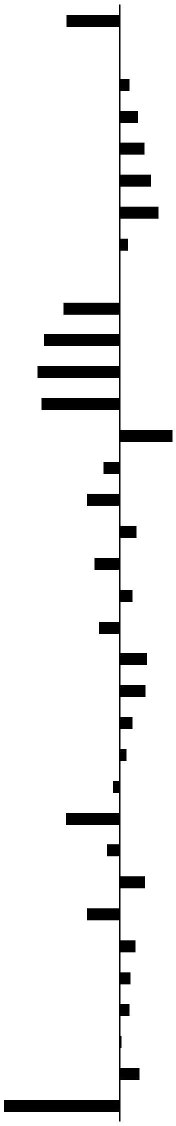

Table 3: The coefficients learned by the LogReg model. The model estimates risk for a given patient profile $\boldsymbol{x} = [x_1 x_2 \ldots x_{32}]$, $x_i \in \{0, 1\}$, as $y(\boldsymbol{x}) = 1/(1 + e^{-z})$, where $z = \beta_0 + \sum_{i=1}^{32} \beta_i x_i$. The MumAge and Parity variables use a 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.

Note that underreporting of maternal conditions leads to inconsistencies in regression coefficients. For example, GRAND MULTIPARITY, which signifies that a mother has had 4 or more children, has a regression coefficient of opposite sign to the coefficient for Parity 4+. In theory, both predictor variables and their coefficients, should be identical. This conflict shows inaccuracies are present in the data recording process.

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 3.

Ensembles of predictors were used in an attempt to moderate over-confident risk estimates from individual predictors [19, 20]. In the absence of any theoretically optimal method to choose the ensemble size, we used Rasmussen's [21] approach of limiting the amount of CPU time used to train any given ensemble.

To train individual predictors 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 predictor 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 [21] and was used instead of more sophisticated (and computationally intensive) second-order methods because of the large size of the SMR2 dataset.

Predictions of individual ensemble members were combined by averaging in the log odds domain [22]. Combining predictions in this way makes the assumption that the log odds of the individual predictors are statistically independent and normally distributed with equal variance. For a given input $\boldsymbol{x}$ and an ensemble of N predictors, the ensemble prediction, $\overline{y}$, is given by

$$\overline{y} = \frac{1}{1 + e^{-\overline{z}}}, \tag{1}$$

where

$$\overline{z} = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{2}$$

and z_i is the log odds of the i^{th} predictor

$$\begin{aligned} z_i &= \log \frac{P_i(\text{adverse}|\boldsymbol{x})}{P_i(\text{benign}|\boldsymbol{x})} \\ &= \log \frac{y_i(\boldsymbol{x})}{1 - y_i(\boldsymbol{x})}. \end{aligned} \tag{3}$$

Figure 4 shows the structure of the networks used in the NetEns and DirEns models. In the networks used in the DirEns model, direct connections between input nodes and the output node were employed to facilitate learning of linear relationships between predictor variables and outcome. As with ensemble size, the number of hidden units used in the neural networks was chosen arbitrarily. Our intention was to use half as many hidden units as there were inputs (14 hidden units) in the first instance and experiment with different architectures if the validation set performance of this configuration proved unsatisfactory.

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

$$P(\text{adverse}|\boldsymbol{x}) \quad = \quad \frac{a_i + \alpha m}{a_i + b_i + \alpha}. \tag{4}$$

Thus, when we have seen many cases corresponding to the profile $\boldsymbol{x}$, this estimate will be dominated by the likelihood $a_i/(a_i + b_i)$. If, however, few or no

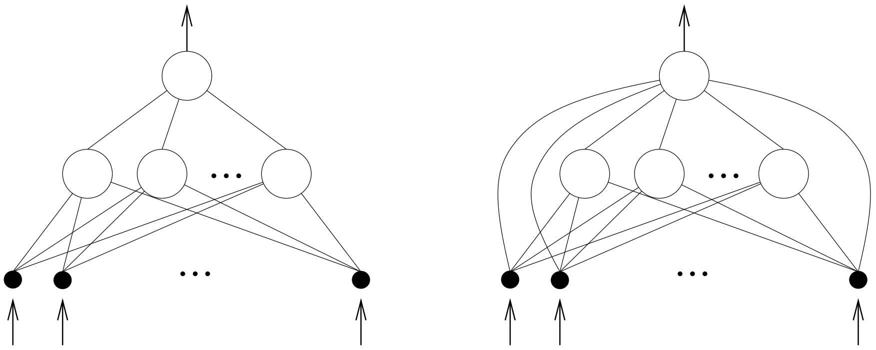

Figure 4: The left diagram shows the structure of the two layer feed-forward networks used in the NetEns prediction model. The diagram on the right shows the kind of fully connected feed-forward network used in the DirEns model.

	All predictors		Non-ICD predictors	
Prediction model	**ANLL**	**AUROC**	**ANLL**	**AUROC**
Test set limits	0.274634	0.7865	0.283611	0.7640
NetEns	0.282474	0.7723	0.284119	0.7634
LkpTab	0.283043	0.7711	0.284128	0.7634
DirEns	0.283527	0.7706	0.284182	0.7632
LogEns	0.283522	0.7702	0.285024	0.7622
LogReg	0.283567	0.7697	0.285031	0.7621

Table 4: The average negative log-likelihood and area under ROC achieved by each of the five prediction models on the test set. The first row shows the minimum attainable ANLL and the MAD for the test set. Columns 2 and 3 show the performances of systems trained using all predictors listed in Table 2. Columns 4 and 5 show the performances of systems trained using only `CSections`, `MumAge`, `NeoDxs` and `Parity`.

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 $\boldsymbol{x}$ 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 [11] and yielded values of $\alpha = 6.012$ and $m = 0.08267$.

3.4 Evaluation of prediction systems

Two statistics were calculated to allow comparison between test set prediction performances of the five models. The *area under the ROC curve* (AUROC) has already been discussed as a way to measure how well a system discriminates between adverse and benign outcomes. The *average negative log-likelihood* (ANLL) of the test data indicates how accurately a model predicts the probability of adverse outcome observed in the test data:

$$\begin{aligned} \text{ANLL} &= \frac{1}{N} \sum_{\boldsymbol{x}_i \in X} a_i \log P(\text{adverse}|\boldsymbol{x}_i) + b_i \log P(\text{benign}|\boldsymbol{x}_i) \\ &= \frac{1}{N} \sum_{\boldsymbol{x}_i \in X} a_i \log y(\boldsymbol{x}_i) + b_i \log(1 - y(\boldsymbol{x}_i)), \end{aligned} \tag{5}$$

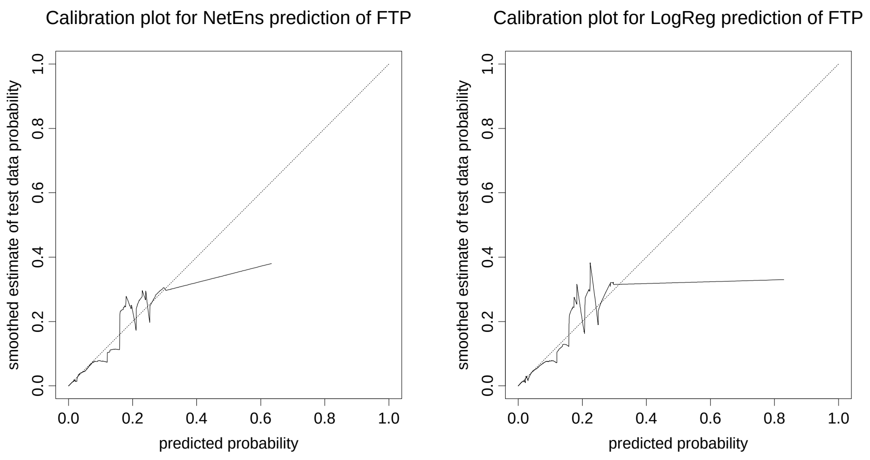

Figure 5: Probability calibration plots for the test set predictions of the NetEns and LogReg models.

where X is the set of distinct patient profiles occurring in the test data, a_i and b_i are the numbers of adverse and benign outcomes for profile $\boldsymbol{x}_i$, 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

$$y(\boldsymbol{x}_i) \quad = \quad \frac{a_i}{a_i + b_i}. \tag{6}$$

Table 4 shows the minimum ANLL, the MAD, and the performances of the five models on the 176812 records in the test set.

Probability calibration plots [23] give a qualitative indication of how well predicted risk corresponds to the risk actually observed. Note that patient profiles in which only one case was recorded lead to observed risks of 1.0 or 0.0, so some form of smoothing is required to extract the underlying relationship between observed and predicted risk. Such *ad hoc* adjustment of the data means that the calibration plots of Figure 5 should not be used for model comparison, but they are useful to check that a model is not seriously under- or over-estimating risk in test data.

4 Discussion

The results in Table 4 show that all five risk prediction models attain close to limit of achievable performance on the test set. In other words, our ability to predict failure to progress is limited by lack of discriminative information in the data, rather than shortcomings in the prediction models. By comparing model performances to the minimum ANLL and MAD, we have removed uncertainty about whether our prediction models were adequate for the task at hand.

There is still, of course, some uncertainty that the BEAD heuristic may have led us to discard an important predictor of adverse outcome. While we think it unlikely that there exist predictor variables within the SMR2 data whose inclusion in the prediction models would significantly improve performance, we cannot discount such a possibility until all possible subsets of predictors have been exhaustively tested. The results presented here strongly suggest that the

Predictor		Errors:	total	partial	omitted
	`MumAge`	4			
	`CSections`	2			
	`NeoDxs`	6			
	`Parity`	4			
ICD-9 code for:	Main condition	105	86	17	2
	Condition 2	94	33	15	46
	Condition 3	64	21	4	39
	Condition 4	33	9	3	21
	Condition 5	8	3	1	4
	Condition 6	5	1	0	4

Table 5: A quality audit of 651 randomly selected SMR2 records was performed in 1991. For the 642 records for which clinical records could be found, this table shows the numbers of errors in the predictors selected in Table 2. The final three columns list the number of ICD-9 codes that were totally incorrect, partially incorrect, or simply not recorded. Note that this audit could not account for errors in the case notes themselves.

more expeditious route to improved performance lies in making more informative clinical observations.

The performance achieved with simple logistic regression shows the relationship between the predictors in Table 2 and the risk of failure to progress to be strongly linear. Also, Figure 5 shows the LogReg model is reasonably well calibrated, at least when $P(\text{adverse}|\boldsymbol{x}) < 0.3$ which covers 97.65% of all cases in the test set (the model *overestimates* risk in the remaining cases). Given the medical community's acceptance and understanding of logistic regression, as well as the simplicity and interpretability of that model, the LogReg system may be the most practical method of predicting failure to progress in the SMR2 data.

There are, however, some significant obstacles to using any of the models described here in a practical setting. First is the fact that the SMR2 data were not collected with the goal of risk prediction in mind. As such, it is unfair to expect the data to be as reliable as data collected in a small study in a single hospital. Inaccuracies in the SMR2 data depend strongly on the kind of information being recorded: measurements (*e.g.*, maternal height) and variables with limited choice (*e.g.*, number of previous pregnancies) tend to be more accurately recorded than items where a wide variety of codes can be used. A quality audit of 651 SMR2 record (see Table 5) showed that ICD-9 codes of maternal condition were particularly prone to error. These errors cause incongruities in the prediction systems derived from the data, as demonstrated by the coefficients of the LogReg model (see Figure 3).

This problem, coupled with uncertainties about the exact diagnostic criteria of ICD-9 codes, prompted us to try to predict failure to progress using non-ICD-9 information only. Columns 4 and 5 of Table 4 show the predictions obtained with systems using this reduced set of predictors compare favourably with those using the full predictor set.

The second practical difficulty with the systems described here is due to the way in which SMR2 data are collected. SMR2 records are in-patient based and completed on discharge (although much of the data is collected and recorded at the time of observation). However, the risk prediction systems we have developed are intended for use prior to the onset of labour: a different clinical setting. The way in which maternal conditions are recorded at discharge will be affected by the outcome of the pregnancy. Pregnancies in which complications arose tend to be reported in greater detail than those that proceeded normally, even though similar maternal conditions may have occurred in both circumstances. Such censoring is probably the reason for the disagreement between the regression coefficients of `GRAND MULTIPARITY` and `Parity` 4+ (Table 3). The incidence of `GRAND MULTIPARITY`, 0.118% (Table 2), is much lower than expected; one may surmise that it has mainly been coded in the presence of another complication of pregnancy or labour.

Tackling these kinds of problems will require fundamental changes to medical data collection methods. We have seen that our ability to predict failure to progress is limited, not by the prediction systems used to model the data, but by the data itself. Clearly, the potential of modern risk prediction techniques is yet to be fully exploited in the medical domain. We hope this paper will motivate clinicians to consider what could be done to obtain data that will permit medical risk to be forecasted more accurately.

5 Conclusions

We have presented a systematic, practical approach to the design and implementation of systems to predict obstetric risk on the basis of a substantial amount of patient data. EAD is a meaningful measure of the discriminative power of a set of predictor variables and we have shown how it can be used to help decide which patient information should be used to model risk of adverse outcome. We have also demonstrated how MAD can determine whether prediction performance is limited by an inherent lack of discriminative information in the predictor variables, or an inadequate prediction model.

Five prediction models were described, ranging in complexity from simple logistic regression to ensembles of neural networks. All systems achieved close to the limit of attainable performance, however, problems of data reliability and uncertain diagnostic criteria restrict the practical potential of these models (and indeed any models that use ICD-9 codes in the SMR2 data).

The reader should bear in mind that the primary aim of this paper is to describe our approach to the design, construction and evaluation of medical risk prediction systems. We do not claim to present a finished system for prediction of failure to progress. Our hope is that medical professionals will see how modern prediction methods can estimate failure to progress *to the extent that the SMR2 data will permit*, and so realise that, in this instance, more informative clinical data needs to be collected for better risk predictions to be made.

Acknowledgements

We are pleased to acknowledge the assistance of Dr Susan Cole of the Information and Statistics Division of the National Health Service in Scotland, for collecting and providing us with access to data from the maternity scheme of the Scottish Morbidity Record system. This work was funded by the Quality Assurance in Maternity Care project, EC Biomed Grant Ref. BIOMED 1 CT93 1113. The first author gratefully acknowledges The British Council, for providing travel funding for him to take up the position of research associate with the QAMC project, and also David MacKay, for many useful comments, suggestions and explanations.

References

[1] J. Kittler, "Mathematical methods of feature selection in pattern recognition," *International Journal of Man-Machine Studies*, vol. 7, pp. 609–637, 1975.

[2] J. M. Van Campenhout, "Topics in measurement selection," in *Handbook of Statistics, Vol. 2* (P. R. Krishnaiah and L. N. Kanal, eds.), pp. 793–803, North-Holland, 1982.

[3] 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.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* Wiley, 1973.

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

[6] 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.

[7] 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.

[8] C. W. Therrien, *Decision, Estimation, and Classification: an Introduction to Pattern Recognition and Related Topics.* Wiley, 1989.

[9] 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.

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

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

[12] J. A. Morgan and J. N. Sonquist, "Problems in the analysis of survey data: and a proposal," *Journal of the American Statistical Association*, vol. 58, pp. 415–434, 1963.

[13] 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.

[14] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

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

[16] 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.

[17] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory.* Wiley, 1994.

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

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

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

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

[22] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural Computation*, vol. 7, pp. 867–888, September 1995.

[23] A. P. Dawid, "Probability forecasting," in *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson, and C. B. Read, eds.), pp. 210–218, New York: Wiley, 1986.