Hum. Reprod. Advance Access published June 24, 2004

DOI: 10.1093/humrep/deh341

The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location

G.Condous[1,4], E.Okaro[1], A.Khalid[1], D.Timmerman[2], C.Lu[3], Y.Zhou[3], S.Van Huffel[3] and T.Bourne[1]

[1]Early Pregnancy, Gynaecological Ultrasound and MAS Unit, St George's Hospital Medical School, London, UK
[2]Department of Obstetrics and Gynaecology, University Hospital Gasthuisberg and [3]Department of Electrical Engineering (ESAT), KU Leuven, Belgium

[4]To whom correspondence should be addressed at: Early Pregnancy, Gynaecological Ultrasound and MAS Unit, St George's Hospital Medical School, Cranmer Terrace, London UK. E-mail: gcondous@hotmail.com

BACKGROUND: The aim of this study was to generate and evaluate new logistic regression models from simple demographic and hormonal data to predict the outcome of pregnancies of unknown location (PULs). METHODS: Data were collected prospectively from 185 consecutive women classified as having a PUL by transvaginal scan; blood was taken at presentation and 48 h later to measure serum progesterone and HCG. These women were followed-up until the outcome was established: an intrauterine pregnancy (IUP), an ectopic pregnancy (EP) or a failing PUL. Three multi-categorical logistic regression models were tested. M1 was based on the HCG ratio (rate of change in HCG over 48 h), M2 was based on the average progesterone level (the mean of the progesterone level at 0 and 48 h) and M3 was based on the patient's age. RESULTS: A total of 102 failing PULs, 63 IUPs and 20 EPs were used in the training set to develop the new models. The best of these models, M3, gave a retrospective area under the receiver operating characteristic (ROC) curve of 0.984 for failing PUL, 0.995 for IUP and 0.920 for EP. All three models were tested prospectively on the test set of 196 cases. M1 outperformed M2 and M3 when tested prospectively. The area under the ROC curve (AUC) was 0.975 for failing PUL, 0.966 for IUP and 0.885 for EP. M1, for the detection of EP, had a sensitivity of 91.7%, a specificity of 84.2%, a positive likelihood ratio of 5.8, a positive predictive value of 27.5% and a negative predictive value of 99.4%. CONCLUSIONS: The logistic regression model M1, can predict which PULs will become failing PULs, IUPs and, most importantly, EPs based on the patient's HCG ratio alone.

Key words: ectopic pregnancy/failing PUL/intrauterine pregnancy/logistic regression/pregnancy of unknown location (PUL)

Introduction

Ten percent of women who present to an Early Pregnancy Unit (EPU) have a pregnancy of unknown location (PUL) (Banerjee *et al.*, 2001). This clinical entity can be defined as a women with a positive pregnancy test with an empty uterus with no signs of an extra-uterine pregnancy on transvaginal ultrasonography. In this situation, there are four possible clinical outcomes: a failing PUL, an intrauterine pregnancy (IUP), an ectopic pregnancy (EP) or a persisting PUL. The location of the failing PUL remains unknown—this group are never visualized using a transvaginal scan (TVS) and an indeterminate proportion of these are failing EPs as well as failing IUPs. Until recently, the small subset of persisting PULs had not been recognized. They are defined as those PULs where the serum HCG levels fail to decline, where there is no evidence of trophoblast disease, and the location of the pregnancy cannot be identified whether by ultrasound or laparoscopy (Condous, 2004). In general, the serum HCG levels are low (<500 IU/l) and have reached a plateau. Prior to this study, we had treated four such women successfully with methotrexate 50 mg/m^2, and their serum HCG levels subsequently resolved (Condous, 2004).

Ten percent of PULs are EPs (Banerjee *et al.*, 2001). EPs account for 80% of early pregnancy deaths (Why Mothers Die, 1997–1999), and therefore the ability to predict whether a PUL is an EP remains a great challenge. Currently the established hormonal criteria for the diagnosis of EP are derived from pregnancies associated with pain and abnormal bleeding and not from asymptomatic women, who will have a much lower pre-test probability of EP.

Previous studies have looked at the use of single variable hormonal models for the prediction of PUL outcome. A serum progesterone of <20 nmol/l predicts failing PUL with a positive predictive value (PPV) of $>95\%$ (Banerjee *et al.*, 2001), and a serum HCG increase of $>66\%$ over 48 h predicts an IUP with a PPV of 96.5% (Condous *et al.*, 2002). Unfortunately, the discriminatory zone and a suboptimally rising serum HCG predict EP with a PPV of only 18.2 and 43.5%,

Human Reproduction © European Society of Human Reproduction and Embryology 2004; all rights reserved

respectively (Condous *et al.*, 2002). To date, there is no hormonal index to predict the outcome of persisting PUL.

In this study, we concentrated on developing baseline multi-categorical logit models that could enable the clinician to distinguish between PULs that are failing PULs, IUPs and EPs based on two blood samples taken 48 h apart.

The aim of this study was to generate and evaluate new logistic regression models based on demographic and hormonal parameters to predict the outcome of PUL. The results are compared with those obtained from established diagnostic criteria for the prediction of failing PUL, IUP and EP in women with PUL.

Materials and methods

Data collection

All women were seen in a single EPU (St George's Hospital, London) from June 2001 to December 2002. This unit was open from 09:00 to 21:00 h 7 days a week. We used a TVS to classify early pregnancies as PUL. PUL was defined with TVS as there being no signs of either an intra- or extra-uterine pregnancy or retained products of conception in a woman with a positive pregnancy test (i.e. HCG >5 IU/l). If the location of the pregnancy could not be ascertained by sonographic images, peripheral blood was taken to measure the levels of serum HCG (World Health Organization, Third International Reference 75/537) and progesterone (Roche Elecsys 2010 Progesterone II test) using automated electrochemiluminescence immunoassays (ECLIAs). These levels were measured 48 h later, according to the protocol.

All scans were reviewed and followed-up by the same primary investigator (G.C.). Exclusion criteria were: (i) the visualization of any evidence of an intrauterine sac; (ii) identification of an adnexal mass thought to be an EP; (iii) those with the presence of heterogeneous, irregular tissues within the uterus thought to be an incomplete miscarriage; and (iv) women who were clinically unstable or demonstrated the presence of a haemoperitoneum on ultrasound scan.

Indications for sonography included lower abdominal pain, with or without vaginal bleeding, poor obstetric history or to determine gestational age.

The study group consisted of 388 consecutive women with a PUL. The data from the first 189 women classified as a PUL collected between June 2001 and February 2002 were taken as the training set. Statistical analysis and building of the logistic regression models were based on this data set. The data from the next 199 PULs recruited between March 2002 and December 2002 were taken as the test set in order to evaluate the performance of the models prospectively.

Data collected included serum hormone levels (serum HCG and progesterone taken at presentation and at 48 h), demographics (age and gestation) and ultrasound features (endometrial thickness, the character of its midline echo and the presence or absence of free fluid in the pouch of Douglas). The women were followed-up until an outcome diagnosis was established: failing PUL, an IUP or an EP. There were four women in the training set and three in the test set who had serum HCG levels that plateaued and no pregnancy was seen at any time. These were classified as persisting PUL and were treated with methotrexate therapy and excluded from the analysis. These were not included for model development and validation because the final outcome was unknown in this subgroup and also because the numbers were so few.

If the initial serum progesterone level was <20 nmol/l, the women were classified as having a failing PUL (Banerjee *et al.*, 2001). Spontaneous resolution of the pregnancy was defined as a decrease in the serum HCG level to <5 IU/l with the disappearance of symptoms. The location of these failing PULs remained unknown. Serum HCG levels were repeated within 7 days to confirm the diagnosis. If the serum HCG rise over the 48 h period was $>66\%$ (Condous *et al.*, 2002), the women were classified as having an IUP and were rescanned 2 weeks later to confirm the diagnosis. Women who did not fall into either category were reviewed every 48 h until a diagnosis was made by sonography. The diagnosis of EP was based upon the positive visualization of an adnexal mass. Ultrasonographic diagnosis of an EP was based on the following grey-scale appearances: (i) an inhomogeneous or inconglomerate mass adjacent to the ovary and moving separate to this—we have called this the 'blob' sign; (ii) a mass with a hyperechoic ring around the gestational sac referred to as the 'bagel' sign; or (iii) a gestational sac with a fetal pole with or without cardiac activity. The diagnosis was confirmed subsequently at laparoscopy with histological confirmation of chorionic villi in the fallopian tube. If an EP was not visualized, but there was a high index of suspicion based on symptomatology, clinical findings and suboptimal rises of serial serum HCG levels, a laparoscopy was performed with or without an evacuation of the uterus.

Data analysis

The data have been pre-processed prior to further analysis. Several variables were created by transformation of the original variables. In particular, the HCG ratio refers to the ratio between the two HCG levels, i.e. serum HCG at 48 h/serum HCG at 0 h, which is more informative than a single HCG level alone. Moreover, it was reported that during early normal gestation, the HCG level 'doubles' every 48 h (Kadar *et al.*, 1981). Thus, intuitively, the use of the HCG ratio should be better than using the single HCG levels. The second transformed variable is the progesterone average, i.e. the mean of the two progesterone levels in an interval of 48 h ([serum progesterone at 0 h + serum progesterone at 48 h]/2). It is accepted that during the period of gestation, progesterone levels rise slightly with time or reach a plateau instead of falling dramatically. Hence the progesterone level at 0 h should be close to that at 48 h. It was also observed that the progesterone level average was distributed extremely dispersedly. The averaged progesterone levels were thus transformed further by taking the logarithm.

Statistical analyses were conducted with SAS (version 8.2 for windows). Univariate and multivariate analysis was performed retrospectively on the basic data (training data) in order to highlight the most significant variables in the model development. To compare the group means for the continuous variables, non-parametric Wilcoxon rank sum tests were used, since most of the continuous variables were not normally distributed. For categorical variables, Fisher's exact tests were used to check their association between the groups. A P-value <0.05 was considered to indicate statistical significance.

Model building

Baseline multi-categorical logit models (Agresti, 1996) were constructed to investigate the relationship between the selected variables and the outcome of the PULs. In such a model, each outcome category is paired with a baseline category, i.e. IUP, resulting in two logit equations, revealing contrasts of the EP versus IUP group and the failing pregnancy versus IUP group. Variables were selected by stepwise procedure with the entry and stay significance level of P-value <0.05.

Performance measure and classification rules

Predictions can be made for the three models by using thresholds (cut-offs) on the output probability of the model. However, the setting of the threshold will influence the accuracy of the prediction. The choice of threshold might vary from institution to institution, and depends on the trade-off between the sensitivity and false-positive rate. In order to see the potential predictive power of those three multi-categorical logit models for each individual category, we firstly considered three binary classification problems, i.e. we used the predicted probability for a certain class to distinguish that class of PULs from the other PULs. Receiver operating characteristic (ROC) analysis can be performed on the three binary classifications independently of class distributions and error costs. The ROC curve for a binary classifier is constructed by plotting the sensitivity (true positive rate) versus 1 – specificity (false-positive rate) for varying cut-off values. The area under the ROC curves (AUC) can be interpreted statistically as the probability of the test correctly distinguishing the abnormal patients from normal ones. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. In this study, the AUC was obtained by a non-parametric method based on the Wilcoxon statistic, using the trapezoidal rule, to approximate the area and its associated standard error (Hanley and McNeil, 1982). This also allowed the comparison of two ROC curves (Hanley and McNeil, 1983).

The performance of the models was also evaluated in terms of sensitivity, specificity, PPV and negative predictive value (NPV).

In order to classify a case into one of the three categories, we needed to set up some diagnostic rules. The rules can be proposed as follows: if the predicted probability for a PUL to be an EP was greater than a threshold, then it was classified as an EP, otherwise it was classified as non-EP. For PULs which were classified as non-EP, if the predicted probability for a PUL to be failing was greater than a threshold, then it was classified as a failing pregnancy; otherwise it was classified as an IUP.

As can be seen from the rules, two probability cut-offs need to be decided. Here we find the 'best' cut-offs by minimizing the square root of $[(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2]$, with the hope that the sensitivity and specificity can both be maximized. The cut-off for EP versus non-EP was based on the predicted probability for a PUL to be an EP given the observation, and the second cut-off for distinguishing failing PULs from IUPs (among non-EPs) was sought using the predicted probability of failing to discriminate between failing and non-failing PULs.

One can form the classification rules based on the weighted predicted probability, which incorporates the probability output of the model with the misclassification cost for different classes. Given an observation, the multi-categorical logistic regression model can provide the predicted posterior probability (P) for each class including P_{ectopic}, P_{failing} and P_{IUP}. We assume that the costs (C) for misclassifying a failing PUL, an IUP or an EP are equal to C_{failing}, C_{IUP} and C_{ectopic}, respectively. For simplicity, here the misclassification costs are assumed to be the same for a certain category of PULs, no matter which class a PUL is wrongly assigned to. The weighted predicted probability for each class can be computed as $C_{\text{failing}}P_{\text{failing}}$, $C_{\text{IUP}}P_{\text{IUP}}$ and $C_{\text{ectopic}}P_{\text{ectopic}}$. The predicted class is then the one with the highest weighted predicted probability among the three. The 'optimal' costs for misclassification were chosen according to the training performance.

Model validation

The models were first validated on the training set by use of ROC analysis for three binary classification problems and the confusion tables for the three-category classification problem. We also utilized the bootstrap technique in order to obtain nearly unbiased estimates of the predictive ability of the models (Efron and Tibshirami, 1993). A total of 100 random samples of the same size as the initial data set were drawn with replacement from the initial data set. Then the logistic models were fitted on each bootstrap sample, and the performance was measured both on the bootstrap sample and on the original sample. The average difference between the two performance measures forms an estimate of the optimism. The bias corrected performance measure was then calculated by subtracting the optimism from the measure of the model built on the original data.

Then the models were validated further on an independent data set with 196 PULs after excluding the persisting PULs. The predicted probabilities of the three classes were calculated with the model developed on the training data, based on which the AUCs and confusion tables were obtained for the test data. Additional bootstrap validation was also performed on the test set.

Results

A total of 388 consecutive women with PUL were recruited, 189 in the training set and 199 in the test set. In the training set of 189, there were 102 (54%) failing PULs, 63 (33.3%) IUPs, 20 (10.6%) EPs and four (2.1%) persisting PULs. The four persisting PULs in the training set were excluded from the further analysis for model building.

In the test set of 199 PUL, there were 109 (54.8%) failing PULs, 75 (37.7%) IUPs, 12 (6.0%) EPs and three (1.5%) persisting PULs. In the test set, 136 (69.4%) presented with lower abdominal pain and 60 (30.6%) without. In the test set, 66 (33.7%) presented without any vaginal bleeding, 68 (34.7%) had vaginal bleeding without clots and 62 (31.6%) had vaginal bleeding with clots.

Table I presents the demographic, hormonal and ultrasonographic characteristics of women with a PUL. Also reported are the P-values for statistical significance of the variables for distinction between groups. These results indicated that almost all the variables seem to be significant in discrimination between an IUP and a non-IUP (either a failing or an EP). On the contrary, none of the variables appeared to be significant for distinguishing an EP from a non-EP (including failing PUL and IUP).

Predictive three-category logistic regression models

Via the stepwise logistic regression, three subsets of variables were selected from different sets of candidate variables, either excluding the log of progesterone average, or excluding the HCG ratio, or including all the variables. The resulting three baseline logit models, namely M1, M2 and M3, were then fitted using the 185 PUL data and are given as follows.

M1 included the HCG ratio alone:

$$\log\left(\frac{P_{\text{ectopic}}}{P_{\text{IUP}}}\right) = 5.79 - 4.21 \ \text{hCG ratio},$$

Table I. Demographic, hormonal and ultrasonographic characteristics of women with a pregnancy of unknown location

Variable	Descriptive statistics for each group				Difference of the variable between groups (*P*-value)				
	Failing PUL	IUP	EP	Persisting PUL	Failing versus IUP	Failing versus EP	IUP versus EP	Failing + IUP versus EP	Failing + EP versus IUP
(a) On training set	$n = 102$	$n = 63$	$n = 20$	$n = 4$					
HCG at 0 h (IU)	594.9 ± 894.1	781.3 ± 1323.3	1510.7 ± 2374	410.8 ± 437.4	0.0046	0.0274	0.7063	0.1182	0.0185
HCG at 48 h	273.7 ± 459.0	1484.9 ± 1958.0	1421 ± 2234	408.5 ± 498.8	<0.0001	0.0001	0.0836	0.0818	<0.001
HCG ratio	0.498 ± 0.333	2.18 ± 0.47	1.068 ± 0.452	0.927 ± 0.254	<0.0001	<0.0001	<0.0001	0.4058	<0.0001
Progesterone at 0 h (IU)	12.16 ± 18.46	68.98 ± 27.68	28.80 ± 19.00	7.75 ± 6.90	<0.0001	<0.0001	<0.0001	0.4218	<0.0001
Progesterone at 48 h	7.54 ± 14.76	68.41 ± 39.88	21.85 ± 13.29	9.0 ± 4.97	<0.0001	<0.0001	<0.0001	0.2121	<0.0001
Progesterone average	9.85 ± 15.89	68.70 ± 32.95	25.33 ± 14.62	8.38 u 5.72	<0.0001	<0.0001	<0.0001	0.3110	<0.0001
Log progesterone average	1.69 ± 0.99	4.13 ± 0.47	3.06 ± 0.63	1.96 ± 0.66	<0.0001	<0.0001	<0.0001	0.3110	<0.0001
Endometrial thickness (mm)	9.87 ± 4.90	14.73 ± 4.53	11.17 ± 6.28	8.43 ± 5.30	<0.0001	0.4844	0.0049	0.4348	<0.0001
Gestational age (days)	48.86 ± 15.43	39.18 ± 14.42	44.20 ± 13.70	40.75 ± 23.04	<0.0001	0.1732	0.0136	0.8632	<0.0001
Age (years)	31.35 ± 6.54	28.64 ± 5.66	29.65 ± 4.15	33.5 ± 5.45	0.0118	0.1701	0.6712	0.4930	0.0216
Disrupted midline echo	16.67%	7.94%	30.00%	0.00%	0.1566	0.2090	0.0201	0.0894	0.0538
Presence of free fluid	9.80%	31.75%	30.00%	0.00%	0.0006	0.0250	1.0000	0.2324	0.0033
(b) On the test set	$n = 109$	$n = 75$	$n = 12$	$n = 3$					
HCG at 0 h (IU)	287.1 ± 457.4	640.1 ± 643.4	567.1 ± 445.6	761.3 ± 614.9					
HCG at 48 h	115.4 ± 177.5	1219.9 ± 986.1	764.5 ± 605.9	1165.7 ± 1167.6					
HCG ratio	0.482 ± 0.459	2.120 ± 0.593	1.403 ± 0.382	1.323 ± 0.400					
Progesterone at 0 h (IUs)	7.78 ± 10.32	71.45 ± 27.49	58.08 ± 39.22	35.67 ± 19.09					
Progesterone at 48 h	4.74 ± 6.78	66.69 ± 29.98	49.33 ± 34.21	30.67 ± 11.85					
Progesterone average	6.26 ± 8.09	69.07 ± 28.01	53.71 ± 35.94	33.17 ± 15.37					
Log progesterone average	1.38 ± 0.88	4.13 ± 0.52	3.69 ± 0.92	3.41 ± 0.58					
Endometrial thickness (mm)	8.51 ± 3.78	13.68 ± 5.50	12.37 ± 3.70	9.83 ± 8.57					
Gestational age (days)	45.47 ± 10.70	39.81 ± 12.01	53.78 ± 21.80	27.50 ± 14.85					
Age (years)	29.27 ± 6.39	29.20 ± 5.36	32.00 ± 4.71	29.67 ± 8.51					
Disrupted midline echo	16.60%	0.00%	25.00%	0.00%					
Presence of free fluid	9.17%	33.33%	25.00%	66.67%					

Data are expressed as means ± SD, or numbers and percentages.

$$\log\left(\frac{P_{\text{failing}}}{P_{\text{IUP}}}\right) = 9.92 - 7.66 \text{ hCG ratio}.$$

We can also get:

$$\log\left(\frac{P_{\text{ectopic}}}{P_{\text{failing}}}\right) = -4.12 + 3.45 \text{ hCG ratio}.$$

P_{ectopic}, P_{failing} and P_{IUP} denote the probability for a patient having an EP, failing PUL or an IUP, respectively. The model can be written equivalently in the following way, by which the predicted probability for each class can be calculated directly:

$$P_{\text{ectopic}} = \frac{e^{5.79-4.21 \text{ hCG ratio}}}{1 + e^{5.79-4.21 \text{ hCG ratio}} + e^{9.92-7.66 \text{ hCG ratio}}},$$

$$P_{\text{failing}} = \frac{e^{9.92-7.66 \text{ hCG ratio}}}{1 + e^{5.79-4.21 \text{ hCG ratio}} + e^{9.92-7.66 \text{ hCG ratio}}},$$

$$P_{\text{IUP}} = \frac{1}{1 + e^{5.79-4.21 \text{ hCG ratio}} + e^{9.92-7.66 \text{ hCG ratio}}}.$$

M2 includes the logarithm (log) of the progesteron average alone:

$$\log\left(\frac{P_{\text{ectopic}}}{P_{\text{IUP}}}\right) = 7.67 - 2.37 \text{ log Prog average},$$

$$\log\left(\frac{P_{\text{failing}}}{P_{\text{IUP}}}\right) = 12.63 - 3.78 \text{ log Prog average}.$$

$$P_{\text{ectopic}} = \frac{e^{7.67-2.37 \text{ log Prog average}}}{1 + e^{7.67-2.37 \text{ log Prog average}} + e^{12.63-3.78 \text{ log Prog average}}},$$

$$P_{\text{failing}} = \frac{e^{12.63-3.78 \text{ log Prog average}}}{1 + e^{7.67-2.37 \text{ log Prog average}} + e^{12.63-3.78 \text{ log Prog average}}},$$

$$P_{\text{IUP}} = \frac{1}{1 + e^{7.67-2.37 \text{ log Prog average}} + e^{12.63-3.78 \text{ log Prog average}}}.$$

M3 utilizes three variables, i.e. hCG ratio, log of progesterone average and age:

$$\log\left(\frac{P_{\text{ectopic}}}{P_{\text{IUP}}}\right) = 12.31 - 4.73 \text{ hCG ratio} - 2.31 \text{ log Prog average} + 0.09 \text{ age},$$

$$\log\left(\frac{P_{\text{failing}}}{P_{\text{IUP}}}\right) = 14.68 - 7.65 \text{ hCG ratio} - 3.63 \text{ log Prog average} + 0.24 \text{ age}.$$

We can also get:

$$\log\left(\frac{P_{\text{ectopic}}}{P_{\text{failing}}}\right) = -2.37 + 2.92 \text{ hCG ratio} + 1.33 \text{ log Prog average} - 0.15 \text{ age}.$$

Thus the predicted probability for a woman having an EP, a failing PUL or an IUP can be obtained from M3 by the following equations:

$$P_{\text{ectopic}} = \frac{e^{12.31-4.73 \text{ hCG ratio}-2.31 \text{ log Prog average}+0.09 \text{ age}}}{1 + e^{12.31-4.73 \text{ hCG ratio}-2.31 \text{ log Prog average}+0.09 \text{ age}} + e^{14.68-7.65 \text{ hCG ratio}-3.63 \text{ log Prog average}+0.24 \text{ age}}},$$

$$P_{\text{failing}} = \frac{e^{14.68-7.65 \text{ hCG ratio}-3.63 \text{ log Prog average}+0.24 \text{ age}}}{1 + e^{12.31-4.73 \text{ hCG ratio}-2.31 \text{ log Prog average}+0.09 \text{ age}} + e^{14.68-7.65 \text{ hCG ratio}-3.63 \text{ log Prog average}+0.24 \text{ age}}}$$

$$P_{\text{IUP}} = \frac{1}{1 + e^{12.31-4.73 \text{ hCG ratio}-2.31 \text{ log Prog average}+0.09 \text{ age}} + e^{14.68-7.65 \text{ hCG ratio}-3.63 \text{ log Prog average}+0.24 \text{ age}}}.$$

We can also get:

$$\log\left(\frac{P_{\text{ectopic}}}{P_{\text{failing}}}\right) = -4.96 + 1.41 \text{ log Prog average}.$$

The predicted probability for each different class can be computed by using these formulae:

All the parameters in the three multi-category logit models are significant, with P-values <0.01, except for the variable age in M3. In model M3, age has a P-value of 0.34 in the equation for the contrast of EP versus IUP group, and a P-value of 0.04 for failing PUL versus IUP group. Table II presents the odds ratios for the HCG ratio, the log

Table II. Multi-category logistic regression analysis with odds ratios (ORs) for the prediction of outcome for the women with a pregnancy of unknown location

Model	Variable	OR (95% CI)	
		EP versus IUP	Failing versus IUP
M1	HCG ratio	0.015 (0.002–0.100)	<0.001 (<0.001–0.005)
M2	Log progesterone average	0.094 (0.032–0.274)	0.023 (0.007–0.073)
M3	HCG ratio	0.009 (<0.001–0.121)	<0.001 (<0.001–0.010)
	Log progesterone average	0.100 (0.023–0.439)	0.026 (0.005–0.133)
	Age	1.094 (0.893–1.340)	1.275 (1.009–1.610)

progesterone average and age, according to the outcome of the pregnancy.

ROC analysis on the three binary classification problems

In order to evaluate the model, we performed ROC analysis for prediction of each outcome category by combining the remaining two categories, i.e. to study the performances of the model for distinguishing ectopic from other (including failing PUL and intrauterine) pregnancies, failing from other (including ectopic and intrauterine) pregnancies, and intrauterine from other (including ectopic and failing PUL) pregnancies. Table III reports the AUC of the three models when validating on the training and test set, respectively. The results included not only the apparent AUCs obtained based on the predicted output from the model fitted with the original training set, but also the bias corrected AUCs from the bootstrap validation. Although the AUCs obtained from bootstrap validation seem to be more accurate and with a narrower confidence interval (CI), the ranking of the performance is the same for the two validation techniques. Thus only some apparent results (without the use of bootstrap) will be highlighted as follows.

As to the validation performance on the training set, M1 gave an AUC of 0.962 (95% CI 0.936–0.988) for predicting

Table III. Areas under the ROC curves of the three multi-category logit models to distinguish a failing PUL, an intrauterine pregnancy or an ectopic pregnancy from the other types of pregnancies for women with a pregnancy of unknown location

Predict class		On training set		On test set	
		AUC (95% CI)	Bias corrected AUC (95% CI)	AUC (95% CI)	Bias corrected AUC (95% CI)
Failing PUL	M1	0.962 (0.936–0.988)	0.961 (0.943–0.980)	0.975 (0.954–0.997)	0.973 (0.953–1.00)
	M2	0.945 (0.913–0.978)	0.943 (0.917–0.971)	0.986 (0.970–1.00)	0.986 (0.976–1.00)
	M3	0.984 (0.966–1.00)	0.980 (0.965–1.00)	0.974 (0.952–0.996)	0.975 (0.959–0.993)
IUP	M1	0.985 (0.964–1.00)	0.985 (0.972–1.00)	0.966 (0.936–0.995)	0.963 (0.941–0.991)
	M2	0.961 (0.926–0.995)	0.960 (0.941–0.987)	0.955 (0.921–0.989)	0.958 (0.929–0.993)
	M3	0.995 (0.984–1.00)	0.993 (0.982–1.00)	0.967 (0.938–0.996)	0.959 (0.938–0.981)
EP	M1	0.839 (0.728–0.950)	0.823 (0.752–0.929)	0.885 (0.760–1.00)	0.914 (0.842–0.989)
	M2	0.874 (0.772–0.975)	0.867 (0.805–0.939)	0.686 (0.515–0.857)	0.728 (0.578–0.849)
	M3	0.920 (0.836–1.00)	0.902 (0.843–0.951)	0.836 (0.693–0.979)	0.834 (0.756–0.946)

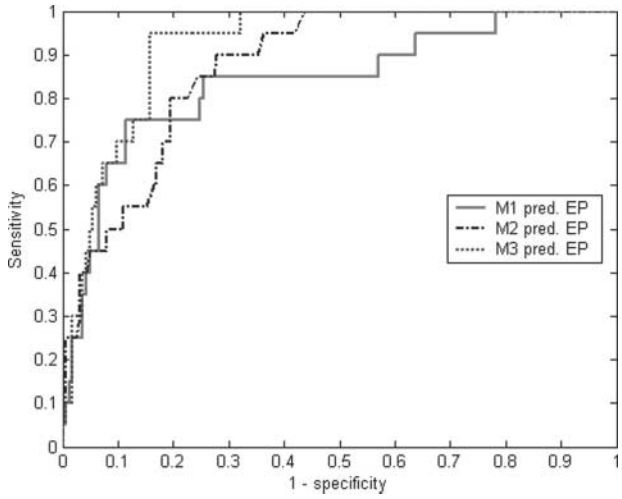


Figure 1. Comparison of ROC curves for prediction of ectopic pregnancy (EP) by using different logit models (on 185 items of training data).

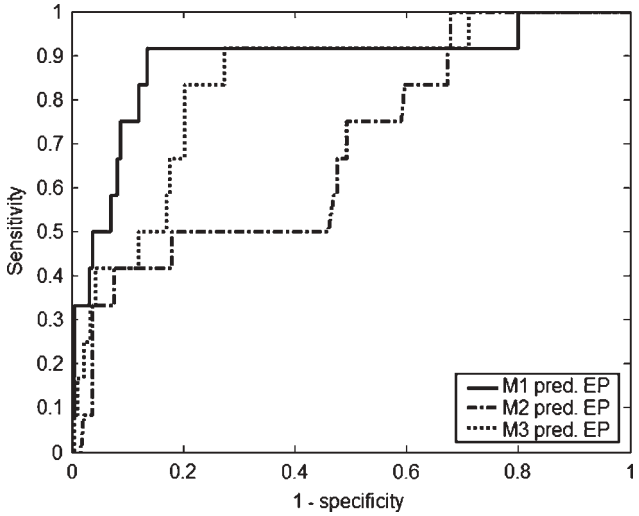


Figure 2. Comparison of ROC curves for prediction of ectopic pregnancy (EP) by using different logit models (on 196 items of test data for M1, and on 195 items of test data for M2 and M3).

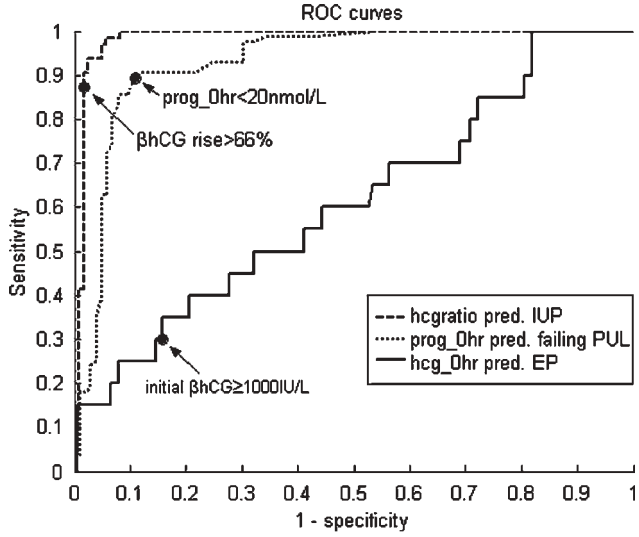


Figure 3. ROC curves for current diagnostic criteria on the training set. AUC for an HCG rise greater than or equal to a cut-off value to predict intrauterine pregnancy (IUP) = 0.985 (95% CI 0.964–1.000); AUC for initial progesterone level less than a cut-off value to predict a failing PUL = 0.930 (95% CI 0.889–0.970); AUC for an initial HCG greater than or equal to a cut-off value to predict ectopic pregnancy (EP) = 0.611 (95% CI, 0.473–0.749).

failing PUL, 0.985 (95% CI 0. 964–1) for IUP, and 0.839 (95% CI 0.728–0.950) for EP. M2 gave an AUC of 0.945 (95% CI 0.913–0.978) for failing PUL, 0.961 (95% CI 0.926–0.995) for IUP, and 0.874 (95% CI 0.772–0.975) for EP. M3 gave an AUC of 0.984 (95% CI 0.966–1) for failing PUL, 0.995 (95% CI 0.984–1) for IUP, and 0.920 (95% CI 0.836–1) for EP. Figure 1 depicts the ROC curves for M1, M2 and M3 for the prediction of EP in the training set.

Although M3 performed the best on the training set, M1 outperformed M2 and M3 when tested prospectively. Since

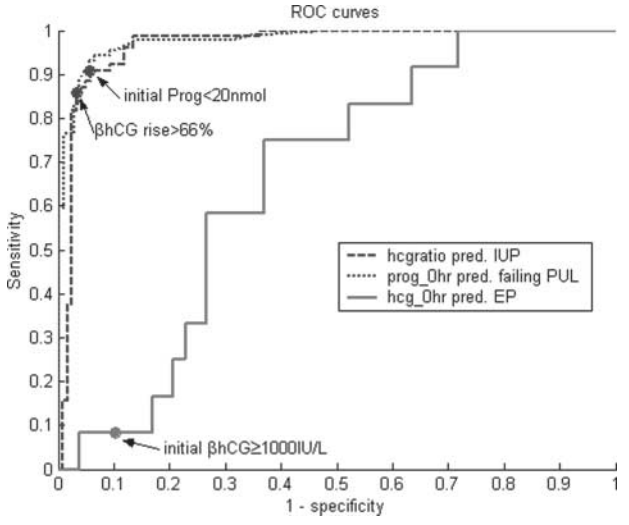


Figure 4. ROC curves for current diagnostic criteria on the test set. AUC for an HCG rise greater than or equal to a cut-off value to predict intrauterine pregnancy (IUP) = 0.966 (95% CI 0.936–0.995); AUC for initial progesterone level less than a cut-off value to predict a failing PUL = 0.980 (95% CI 0.959–1.000); AUC for an initial HCG greater than or equal to a cut-off value to predict ectopic pregnancy (EP) = 0.666 (95% CI 0.494–0.839).

one woman with a failing PUL in the test set had no measurement of progesterone level, only 195 test PULs were used for M2 and M3, while all the 196 test PULs were used for M1. M1 gave an AUC curve of 0.975 (95% CI 0.954–0.997) for failing PUL, 0.966 (95% CI 0.936–0.995) for IUP, and 0.885 (95% CI 0.760–1) for EP. Figure 2 depicts the ROC curves for M1, M2 and M3 for the prediction of EP in the test set.

For reasons of comparison, ROC curves were constructed in order to illustrate the discriminative capacity of using serum HCG and serum progesterone measurements to diagnose PUL (see Figures 3 and 4). Also shown on the ROC curves are the operating points corresponding to the current diagnostic criteria: serum HCG rise over the 48 h period of >66% to predict an IUP (Agresti, 1996); initial serum progesterone of <20 nmol/l to predict a failing pregnancy (Banerjee *et al.*, 2001); and serum HCG of ≥1000 IU/l at presentation to predict an EP (Hanley and McNeil, 1982).

M1 performed better than the discriminatory zone on both the training and the test set for the prediction of PUL. Using serum HCG at presentation to predict an EP gave an AUC on the test set of 0.666 (95% CI 0.494–0.839) compared with 0.885 (95% CI 0.760–1) given by M1. M1 was comparable with using a serum HCG rise over 48 h to predict IUP (both reach an AUC of 0.966 on the test set) and the initial serum progesterone to predict a failing PUL (both reach an AUC of 0.980; 0.975 on the test set).

Classification rules for the three-category classification problem

From the best performing model (on the test set), M1, we computed the posterior probability for a woman having a failing PUL, an IUP or an EP.

Since model M1 has only one explanatory variable, the HCG ratio, we can visualize the relationship between the posterior probability and HCG ratio, as shown in Figures 5 and 6. The dotted line indicates the predicted probability for an observation being an ectopic PUL versus its HCG ratio, the solid line for an observation being a failing PUL, and the dashed line for an observation being an IUP. Also shown in the figure is the observed probability of a PUL being a failing PUL, EP or IUP given the HCG ratio. The variable HCG ratio was first divided into 12 evenly spaced intervals between 0 and 4, then the observed probability was estimated by the proportion of an outcome category within each interval using the data from the training and test set, respectively. The predicted and observed probabilities seem to match quite well in both the training and test data. There is only one exceptional extreme case in the test set when the HCG ratio is close to 4 (see Figure 6). However, the observed probability is not reliable for this last interval with the HCG ratio >3.64, since there is only one PUL (a failing PUL from the test set) in this interval.

For the three-category classification, we first checked the performance of the model using the following rules. If the predicted probability for a PUL to be an EP was >0.21, then it was classified as an EP, otherwise it was classified as an a non-EP. For a PUL which was classified as a non-EP;

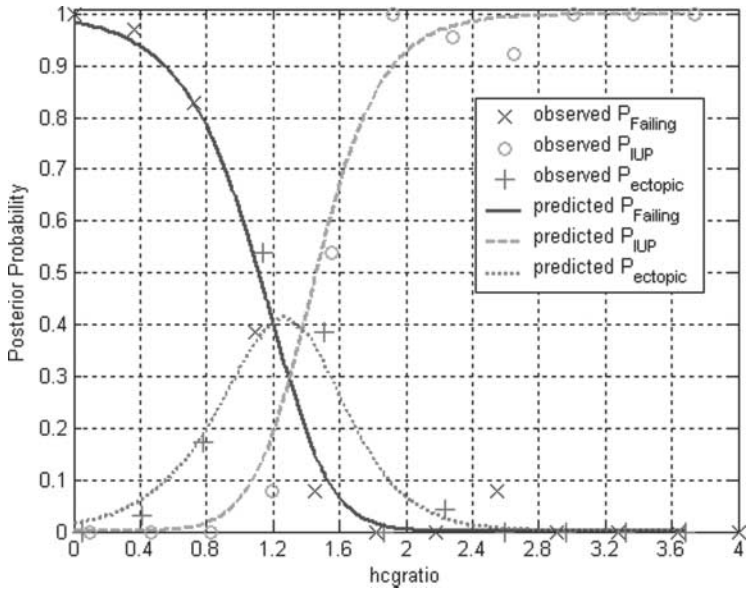


Figure 5. Posterior probabilities for a woman having a failing PUL, an intrauterine pregnancy (IUP) or an ectopic pregnancy (EP) versus the HCG ratio. The predicted probability was given by model M1, and the observed probability was calculated from the training set.

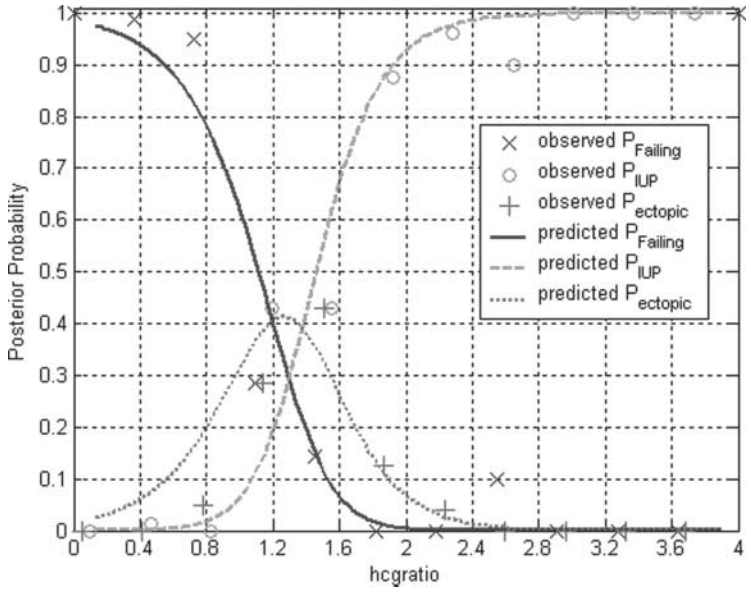


Figure 6. Posterior probabilities for a woman having a failing PUL, an intrauterine pregnancy (IUP) or an ectopic pregnancy (EP) versus the HCG ratio. The predicted probability was given by model M1, and the observed probability was calculated from the test set.

Table IV. The performance of model M1 when using the classification rules with two probability cut-off values

	Training					Test			
	Failing	IUP	Ectopic	Total		Failing	IUP	Ectopic	Total
Failing	91	1	10	102	Failing	98	2	9	109
IUP	0	54	9	63	IUP	1	61	13	75
Ectopic	4	1	15	20	Ectopic	1	1	10	12
Total	95	56	34	185	Total	100	64	32	196
	Sensitivity	Specificity	PPV	NPV		Sensitivity	Specificity	PPV	NPV
Failing	89.2%	95.2%	95.8%	87.8%	Failing	89.9%	97.7%	98.0%	88.5%
IUP	85.7%	98.4%	96.4%	93.0%	IUP	81.3%	97.5%	95.3%	89.4%
Ectopic	75.0%	88.5%	44.1%	96.7%	Ectopic	83.3%	88.0%	31.3%	98.8%

[a] The cut-off was set to be 0.21 for distinguishing ectopic pregnancies (EPs) from non-EPs, and 0.72 for distinguishing failing PULs from intrauterine pregnancies (IUPs) among the non-EPs

Table V. The performance of model M1 based on the weighted probabilities

(a) The results when setting the cost values to be 1, 1 and 4 for misclassifying a failing PUL, an intrauterine pregnancy (IUP) and an ectopic pregnancy (EP), respectively

	Training				Test			
	Failing	IUP	Ectopic	Total	Failing	IUP	Ectopic	Total
Failing	90	1	11	102	95	2	12	109
IUP	0	53	10	63	1	61	13	77
Ectopic	4	1	15	20	1	1	10	12
Total	94	55	36	185	97	64	35	196
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
Failing	88.2%	95.2%	95.7%	86.8%	87.2%	97.7%	97.9%	85.9%
IUP	84.1%	98.4%	96.4%	92.3%	81.3%	97.5%	95.3%	89.4%
Ectopic	75.0%	87.3%	41.7%	96.6%	83.3%	86.4%	28.6%	98.8%

(b) The results when setting the cost values to be 1, 1 and 5 for misclassifying a failing PUL, an IUP and an EP, respectively

	Training				Test			
	Failing	IUP	Ectopic	Total		Failing	IUP	Ectopic	Total
Failing	88	1	13	102	Failing	94	2	13	109
IUP	0	51	12	63	IUP	1	58	16	75
Ectopic	4	1	15	20	Ectopic	0	1	11	12
Total	92	53	40	185	Total	95	61	40	196
	Sensitivity	Specificity	PPV	NPV		Sensitivity	Specificity	PPV	NPV
Failing	86.3%	95.2%	95.7%	85.0%	Failing	86.2%	98.9%	99.0%	85.2%
IUP	81.0%	98.4%	96.2%	90.9%	IUP	77.3%	97.5%	95.1%	87.2%
Ectopic	75.0%	84.9%	37.5%	96.6%	Ectopic	91.7%	84.2%	27.5%	99.4%

Table VI. The HCG ratio of pregnancies with unknown location (on the whole 381 data): occurrences, predicted probabilities of model M1, and the likelihood ratios for different outcomes of pregnancies

	Occurrence			Probability			Likelihood ratio		
HCG ratio	Failing	IUP	EP	Failing	IUP	EP	Failing	IUP	EP
<0.5	140	1	2	0.98–0.92	0	0.02–0.08	37.60	0.01	0.15
0.5–0.69	38	0	2	0.92–0.84	0–0.01	0.08–0.15	15.31	0.00	0.57
0.7-0.79	9	0	1	0.84–0.79	0.01–0.02	0.15–0.19	7.25	0.00	1.21
0.8–0.89	9	0	3	0.78–0.72	0.02–0.03	0.20–0.25	2.42	0.00	3.64
0.9–0.99	4	2	3	0.71–0.63	0.03–0.06	0.26–0.31	0.64	0.50	5.45
1.0–1.19	5	4	6	0.62–0.41	0.07–0.18	0.32–0.40	0.40	0.64	7.27
1.2–1.39	1	1	4	0.40–0.20	0.19–0.41	0.41–0.39	0.16	0.35	21.81
1.4–1.59	2	7	6	0.19–0.07	0.42–0.66	0.39–0.27	0.12	1.54	7.27
1.6–.65	0	4	2	0.07–0.05	0.67–0.72	0.26–0.23	0.00	3.52	5.45
1.66–1.79	0	13	1	0.05–0.02	0.73–0.83	0.22–0.15	0.00	22.89	0.84
≥ 1.8	3	106	2	≤ 0.02	≤ 0.84	≤ 0.14	0.02	37.33	0.20
Total	211	138	32						

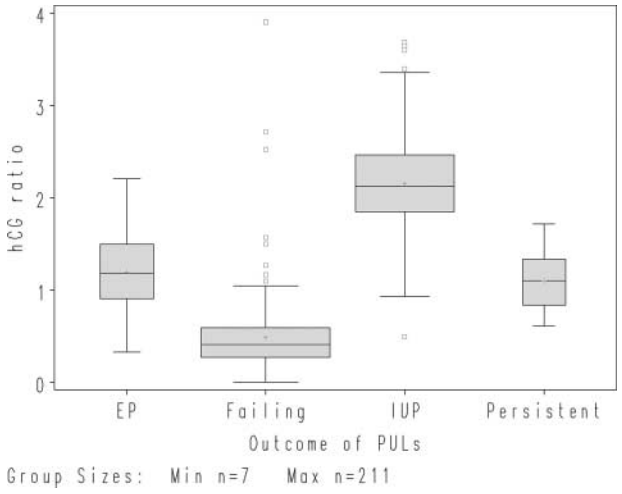


Figure 7. Boxplot of the HCG ratio versus outcome of PULs (on data from all 388 women).

if the predicted probability for this PUL to be failing was >0.72, then it was classified as a failing PUL; otherwise it was classified as an IUP. The two cut-off points were sought out by maximizing both sensitivity and specificity on the training set. The classification results on both training and test set are shown in Table I.

We also tried to derive the classification rules using weighted predicted probabilities, by which we explicitly incorporated the misclassification costs into our decision making. By varying the costs, we obtained different results. The optimal (relative) cost values for misclassifying a failing PUL, an IUP and an EP were 1, 1 and 4, respectively, which were selected based on the performance on the training set. The corresponding results of these rules on the training and test set are shown in Table V(a).

One can see that from Table V(b), if the relative cost for misclassification of an EP is increased to 5 while keeping

the other cost values as 1, M1 achieved a 91.7% detection rate for EP and a false-positive rate of 15.8%.

We also notice that the PPVs for EP are quite low in Tables IV and V, which is due mainly to the low prevalence of EPs in our study group. On the contrary, the likelihood ratios (LRs) are mathematically independent of the prevalence and considered more informative for clinical practice. Therefore, in Table VI, we present the diagnostic results in LR for different intervals of the HCG ratio, together with the occurrence and corresponding ranges of predicted probabilities from M1 for different types of PULs. Since the number of EPs is very small in both the training and the test set, the data from all 381 women have been used in order to obtained a more reliable LR estimates.

Prediction on the persistent PULs

Remember that the persistent PULs were excluded from model development and validation, which might have an impact on the model performance when used in practice. In order to discover the potential influence of the model prediction on these patients, we checked the predicted probability of the seven persistent PULs using model M1. From the boxplot of the HCG ratio of the data from all 388 women, we can see that the distribution of the seven persistent PULs is close to the group of 32 EPs (see Figure 7). Not surprisingly, six out of the seven persistent PULs were predicted to be an EP and one was predicted be a failing pregnancy, when the same classification rules in Table V(b) were exploited based on the weighted probability. Almost the same predictions were given by utilizing the rules in Tables VI and V(a), except that one PUL was predicted to be an IUP instead of an EP. A similar phenomenon can be observed for the prediction of model M2 and M3: at least five persistent PULs were predicted to be an EP. Therefore, it can be inferred that the persistent PULs have similar characteristics to EPs, and a persistent PUL is assigned to an EP by the multi-category logit models with a chance of $>70\%$.

Discussion

Between 87 and 93% of EPs (Cacciatore *et al.*, 1990; Shalev *et al.*, 1998; Rosello *et al.*, 2003) will be visualized on TVS, and this group of women is relatively straightforward to manage as a positive diagnosis has been made. A positive diagnosis of an EP permits a greater range of management options to be considered. When it is not possible to visualize a pregnancy on TVS either inside or outside the endometrial cavity, no firm diagnosis can be made and these women are labelled as having a PUL. These may account for 10% of pregnancies attending an EPU (Banerjee *et al.*, 2001), although the number of PULs will vary according to the quality of scanning in a particular unit. The use and interpretation of serum levels of HCG and progesterone is essential if attempts are to be made to distinguish between failing PUL, IUP and EP in the PUL population. The prediction of the outcome in those women who have a PUL is potentially the most critical function of an EPU. Traditionally, the discriminatory zone (Cacciatore *et al.*, 1990) has been relied upon to select women at increased risk of EP, whilst changing levels of serum HCG (Dart *et al.*, 1999) and the initial serum progesterone (Banerjee *et al.*, 2001) have been used to predict the outcome. We have demonstrated in previous studies that the predictive value of hormonal tests in failing PUL and viable pregnancies is high (Condous *et al.*, 2002). PULs can thus be classified as being at low and high risk of EP. However, those PULs which are an early EP are poorly predicted (Condous, 2004). It is the latter group that is at the highest risk of haemorrhage and death, and yet, to date, the use of hormonal indices has yielded discouraging results. This high-risk group of PULs accounts for only 10% of all PULs (Banerjee *et al.*, 2001), yet their misclassification can lead to serious complications. Current algorithms for the diagnosis of an EP do not take into account the heterogenous nature of the cohort of women presenting with a risk of having this condition. Such heterogeneity can lead to differences in pre-test probability of an EP (Mol *et al.*, 1999a). In women with clinical symptoms, for example, the probability of an EP is higher than in symptom-free women. Any additional tests should be interpreted differently, depending on the pre-test probability. With this in mind, we developed new models that could improve the detection of EPs in this group of women with a PUL, without resulting in unnecessary anxiety, i.e. a high false-positive rate.

In developing the models, gestational age and endometrial thickness were not selected by the stepwise logistic regression procedure, though they appeared to be significant in the univariate logistic regression models. This is due to the strong correlations between these variables and the HCG ratio, which are both ~ 0.4, and their correlations with the average progesterone levels, which are both ~ 0.3. In previous studies, the use of these parameters as an alternative to serum HCG for the diagnosis of EP in women with PUL has not proved diagnostic (Mol *et al.*, 1999b).

The models do not have to be used at the same gestational period in early pregnancy provided the serum HCG levels are $<10\,000$ IU/l. We know that when the serum HCG concentrations are $<10\,000$ IU/l, the rate of serum HCG change does not increase significantly, i.e. it is linear (Kadar *et al.,* 1990). All the women in this study with an early IUP had an initial serum HCG $<10\,000$ IU/l; therefore, the rate of change of the log HCG was linear. Thus the models do not require the same gestational age in PULs.

The performance results of the training sets of all three multi-categorical logistic regression models, M1, M2 and M3, were very encouraging. All three models outperformed current diagnostic criteria for EP and were as good as current diagnostic criteria for predicting viability.

When the AUC in each model (for the prediction of EP on the training set) is compared with single parameters such as the discriminatory zone, we see that their performance is significantly better.

As these results were obtained retrospectively, we needed to cross-validate the results prospectively in order to assess how robustly each model performed. Each model when tested in this way gave equally encouraging results.

One limitation of this study is that the sample size is rather small with regard to the number of events per variable (EPV). This will influence the stability of the stepwise logistic regression, for example the selected variables might change when deleting or adding a small amount of data. Focusing on the prediction of EPs, the EPV values for model M1 and M2 are both 20, which are probably large enough to obtain a stable parameter estimate of the logit models. However, the EPV value is only 6.7 for M3. Moreover, the incidence of the outcome is different between the training set and test set, and the effects of the predictors may also be different. This might partially explain why the validation results from the test set did not agree well with those from the training set. As a post analysis, we combined the training and test sets into one data set for model development. Again we started the stepwise selection from three sets of candidate variables. Both the HCG ratio and the log progesterone average still keep their important roles in the models. Whereas age was ruled out of the models, contradictorily, the disrupted midline echo appeared to be significant in all the three final models. Therefore, a larger data set is still needed in order to develop a more stable model. Based on the current available data, M1 is the most impressive model among the three. It is simple, while its validation performance is still comparable with or even better than the other two models.

Our optimal logistic regression model M1 represents a significant improvement on current diagnostic criteria for the detection of EP. We believe that multi-centre trials are needed to test its reproducibility and validity before it is adopted in the clinical setting. In the future, we hope that the incorporation of historical factors, such as previous history of pelvic inflammatory disease or EP, and the presence or absence of site-specific tenderness at the time of TVS will result in a more detailed modelling of the probability of an EP. In turn, this could result in a better diagnostic performance.

This logistic regression model can predict which PULs become failing PULs, IUPs and, most importantly, EPs based on the patient's HCG ratio alone. It significantly outperforms current diagnostic criteria for the prediction of EPs.

Acknowledgements

This research was supported by interdisciplinary research grants of the research council of the Katholieke Universiteit Leuven, Belgium (IDO/99/03 and IDO/02/009 projects), by the Belgian Programme on Interuniversity Poles of Attraction (IUAP Phase V-22) and the Concerted Action Project MEFISTO-666 of the Flemish Community. C.L. is supported by a KU Leuven PhD scholarship.

References

Agresti A (1996) An Introduction to Categorical Data Analysis. Wiley & Sons, New York.

Banerjee S, Aslam N, Woelfer B, Lawrence A, Elson J and Jurkovic D (2001) Expectant management of early pregnancies of unknown location: a prospective evaluation of methods to predict spontaneous resolution of pregnancy. Br J Obstet Gynaecol 108,158–163.

Cacciatore B, Stenman UH and Ylostalo P (1990) Diagnosis of ectopic pregnancy by vaginal ultrasonography in combination with a discriminatory serum hCG level of 1000 IU/L (IRP). Br J Obstet Gynaecol 10,904–908.

Condous GS (2004) The management of early pregnancy complications. In Bourne T and Valentin L (eds), Best Practice and Research Clinical Obstetrics and Gynaecology Special Issue: Volume 18 Issue 1. Ultrasound in Gynaecology. Elsevier, Amsterdam 37–57.

Condous G, Okaro E, Khalid A, Zhou Y, Lu C, Van Huffel S, Timmerman D and Bourne T (2002) Role of biochemical and ultrasonographic indices in the management of pregnancies of unknown location. Ultrasound Obstet Gynaecol 20 Suppl 1,36–37.

Dart RG, Mitterando J and Dart LM (1999) Rate of change of serial beta-human chorionic gonadotropin values as a predictor of ectopic pregnancy in patients with indeterminate transvaginal ultrasound findings. Ann Emerg Med 34,703–710.

Efron B and Tibshirami RJ (1993) An Introduction to the Bootstrap. Chapman & Hall, New York.

Hanley JA and McNeil B (1982) The meaning and use of the area under a receiver operating characteristic curve. Diagn Radiol 143,29–36.

Hanley JA and McNeil B (1983) A method of comparing the areas under the receiver operating characteristics curves derived from the same cases. Radiology 148,839–843.

Kadar N, Caldwell BV and Romero R (1981) A method of screening for ectopic pregnancy and its indications. Obstet Gynecol 58,162–166.

Kadar N, Freedman, M and Zacher M (1990) Further observations on the doubling time of human chorionic gonadotropin in early asymptomatic pregnancies. Fertil Steril 54 783–787..

Mol BW, van Der Veen F and Bossuyt PM (1999a) Implementation of probabilistic decision rules improves the predictive values of algorithms in the diagnostic management of ectopic pregnancy. Hum Reprod 14, 2855–2862.

Mol BW, Hajenus PJ, Engelsbel S, Ankum WM, van der Veen F, Hemrika DJ and Bossuyt PM (1999b) Are gestational age and endometrial thickness alternatives for serum human chorionic gonadotropin as criteria for the diagnosis of ectopic pregnancy? Fertil Steril 72,643–645.

Rosello N, Condous G, Okaro E, Khalid A, Alkatib M, Rao S and Bourne T (2003) Does transvaginal ultrasonography accurately diagnose ectopic pregnancy? Hum Reprod 18 Suppl 1,160.

Shalev E, Yarom I, Bustan M, Weiner E and Ben-Shlomo I (1998) Transvaginal sonography as the ultimate diagnostic tool for the management of ectopic pregnancy: experience with 840 cases. Fertil Steril 69,62–65.

'Why Mothers Die' Triennial Report 1997-1999. Confidential Enquiry into Maternal Deaths, UK.

Submitted on October 31, 2003; accepted on May 6, 2004