Human Reproduction, Vol.25, No.10 pp. 2434–2440, 2010
Advanced Access publication on August 17, 2010 doi:10.1093/humrep/deq217

human reproduction

ORIGINAL ARTICLE *Early pregnancy*

Does a prediction model for pregnancy of unknown location developed in the UK validate on a US population?

K.T. Barnhart[1,2,*], M.D. Sammel[2], D. Appleby[2], M. Rausch[1], T. Molinaro[1], B. Van Calster[3], E. Kirk[4], G. Condous[5], S. Van Huffel[3], D. Timmerman[6], and T. Bourne[6,7]

[1]Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, University of Pennsylvania, 3701 Market Street, Suite 800, Philadelphia, PA 19104, USA [2]Center for Clinical Epidemiology and Biostatistics, Division of Reproductive Endocrinology and Infertility, Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA [3]Department of Electrical Engineering (ESAT-SISTA), K.U. Leuven, Leuven, Belgium [4]Early Pregnancy and Gynaecological Unit, St George's University of London, Cranmer Terrace, London SW17 0RE, UK [5]Early Pregnancy and Advanced Endosurgery Unit, Nepean Clinical School, University of Sydney, Nepean Hospital, Sydney, Australia [6]Department of Obstetrics and Gynaecology, University Hospital Gasthuisberg, K.U. Leuven, Leuven, Belgium [7]Imperial College London, Hammersmith Campus, Du Cane Road, London W12 0HS, UK

*Correspondence address. E-mail: kbarnhart@obgyn.upenn.edu

Submitted on April 27, 2010; resubmitted on July 2, 2010; accepted on July 7, 2010

BACKGROUND: A logistic regression model (M4) was developed in the UK to predict the outcome for women with a pregnancy of unknown location (PUL) based on the initial two human chorionic gonadotrophin (hCG) values, 48 h apart. The purpose of this paper was to assess the utility of this model to predict the outcome for a woman (PUL) in a US population.

METHODS: Diagnostic variables included log-transformed serum hCG average of two measurements, and linear and quadratic hCG ratios. Outcomes modeled were failing PUL, intrauterine pregnancy (IUP) and ectopic pregnancy (EP). This model was applied to a US cohort of 604 women presenting with symptomatic first-trimester pregnancies, who were followed until a definitive diagnosis was made. The model was applied before and after correcting for differences in terminology and diagnostic criteria.

RESULTS: When retrospectively applied to the adjusted US population, the M4 model demonstrated lower areas under the curve compared with the UK population, 0.898 versus 0.988 for failing PUL/spontaneous miscarriage, 0.915 versus 0.981 for IUP and 0.831 versus 0.904 for EP. Whereas the model had 80% sensitivity for EP using UK data, this decreased to 49% for the US data, with similar specificities. Performance only improved slightly (55% sensitivity) when the US population was adjusted to better match the UK diagnostic criteria.

CONCLUSIONS: A logistic regression model based on two hCG values performed with modest decreases in predictive ability in a US cohort for women at risk for EP compared with the original UK population. However, the sensitivity for EP was too low for the model to be used in clinical practice in its present form. Our data illustrate the difficulties of applying algorithms from one center to another, where the definitions of pathology may differ.

Key words: ectopic pregnancy / prediction model / pregnancy of unknown location

Introduction

Women who present with pain and/or bleeding in the first trimester of pregnancy are at risk for ectopic pregnancy (EP). Systematic evaluation of women at risk has dramatically reduced the morbidity and mortality of the disease; nonetheless, ruptured EPs continue to occur. In many cases, this is due to failure of the clinician, or the patient, to recognize the early signs and symptoms of the condition. In other cases, rupture occurs during the time it takes to confirm the diagnosis (Barnhart, 2009). Unruptured EP can generally be diagnosed rapidly and accurately utilizing transvaginal ultrasound in conjunction with a quantitative serum human chorionic gonadotrophin (hCG) measurement (Seeber and Barnhart, 2006; Kirk *et al.*, 2007a; Barnhart, 2009). However, between 10 and 20% of symptomatic women who undergo an ultrasound scan in early pregnancy will have a non-diagnostic scan (i.e. an empty uterus, and no signs of an intrauterine pregnancy (IUP) or extrauterine pregnancy on transvaginal ultrasound examination). This situation is labeled a pregnancy of unknown location (PUL) (Condous *et al.*, 2004a,b; Kirk *et al.*, 2007a). The management of PUL is often a difficult task, with serious potential medical and ethical consequences for misdiagnosis.

A model to predict the ultimate location of a PUL in women at risk for EP based on the first two hCG values would be valuable if it had excellent test characteristics and could be validated. Before any prediction model should affect standard of care, it should be validated in a sample external from that in which it was developed. The consequence of poor validation is misclassification that can result in interruption of a desired ongoing IUP or rupture of an EP. Often a prediction model has poorer test characteristics when it is validated in a separate population, especially in a population distinct from its development.

A series of mathematical models using repeated measures of serum hCG at 0 and 48 h have been developed to predict the outcome for women with a PUL (Condous *et al*., 2004b, 2007a,b; Kirk *et al*., 2006, 2007b) with the latest, called M4, demonstrating a sensitivity of 80% for EP and a receiver-operating characteristic (ROC) curve (AUC) of 0.90 (Condous *et al*., 2007a). However, all such models have limitations, including external validity. The purpose of this paper is to validate M4, a logistic regression model developed in the UK to distinguish EPs and spontaneous miscarriages from ongoing IUPs, in a US population (Condous *et al*., 2007a).

Materials and Methods

Data collection and cleaning

Approval to conduct this study was obtained from the Institutional Review Board of the University of Pennsylvania. A database of all women in the first trimester of pregnancy (positive pregnancy test or history of a missed period) who present with pain and/or bleeding is maintained at the University of Pennsylvania. Data were entered directly into the computerized database by clinical staff caring for the patient. The database collects medical and surgical history, maternal and gestational age, clinical presentation (symptoms like pain and bleeding) and diagnostic tests (ultrasound and hCG). Women were followed in this clinical database until they were definitively diagnosed with an EP or IUP or spontaneous miscarriage. Spontaneous miscarriage was confirmed either by the histopathology of products of conception on suction dilatation and curettage or by the spontaneous decline of hCG level to $\leq$5 IU/l. The presence of an ongoing IUP was confirmed by observing ongoing progression of the pregnancy by ultrasound with visualization of an intrauterine yolk sac, or fetal pole. The diagnosis of EP was confirmed either by the presence of chorionic villi in the fallopian tube, by visualizing an extrauterine gestational sac (with yolk sac or embryonic cardiac activity) with ultrasonography for those treated medically, or by a rise or plateau in hCG level after dilation and evacuation (and no evidence of chorionic villi in the endometrial curettage sample).

Data in this system were also used to subclassify women with a spontaneous miscarriage into a diagnosis of complete miscarriage, incomplete miscarriage, intrauterine fetal demise or anembryonic gestation. Women with an EP were stratified based on diagnostic criteria: diagnosed at surgery, diagnosed with ultrasound or non-visualized (increase in hCG after uterine evacuation).

Data from 1 February 2003 to 30 September 2007 were used to form the US cohort. Where appropriate, missing data and/or questionable values were double-checked against electronic medical records and charts for validation.

This data set was compared with a UK data set covering the time period from 18 July 2003 to 9 October 2004 from the same setting as that on which M4 was developed. Women presenting to the Early Pregnancy Unit (EPU) at St George's Hospital, London, UK, with a positive urinary pregnancy test underwent a transvaginal ultrasound examination for various reasons. Indications included lower abdominal pain, vaginal bleeding, maternal anxiety or confirmation of gestational age. Women were classified as having a PUL if there was no evidence of an IUP or extrauterine pregnancy on transvaginal sonography. This population is partially represented in four manuscripts (Condous *et al*., 2006; Gevaert *et al*., 2006; Bignardi *et al*., 2008; Van Calster *et al*., 2009).

All women were managed expectantly until the final definitive diagnosis was made, i.e. until either the location of the pregnancy was established using transvaginal ultrasound or the serum hCG levels declined to undetectable levels. There were three outcome groups: a failing PUL, an IUP or an EP. The follow-up protocol was as follows. If the serum hCG rise over the 48-h period was >66%, the women were classified initially as having an early IUP and were rescanned 2 weeks later. The presence at follow-up scan of an intrauterine sac surrounded by a brightly echoic ring, situated eccentrically within the endometrial cavity, confirmed the diagnosis. Spontaneous resolution of the pregnancy was defined as a decrease in the serum hCG level to <5 IU/l. The location of these failing PULs remained unknown. Women who did not fall into either category were reviewed every 48 h with serum hCG testing and/or sonography until a diagnosis was made. A diagnosis of EP was made using ultrasound if a mass was seen in the adnexa with echogenicity consistent with an EP; this included an inhomogeneous mass or empty gestational sac, as well as those with a sac containing a yolk sac or fetal pole (Condous *et al*., 2007a; Kirk *et al*., 2007a). If an EP was not visualized using transvaginal ultrasound, 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 evacuation of the uterus. The gold standard for the diagnosis of tubal EP (histological confirmation of villi in the tube) was not applied to all women, because some women with an ultrasound diagnosis of tubal pregnancy were treated medically.

All statistical analysis was carried out using SAS version 9.1 (SAS Institute, Cary, NC, USA).

Model application

Condous *et al*. (Condous *et al*., 2004b, 2007a,b; Kirk *et al*., 2006, 2007b) reported on multiple models developed to distinguish between EPs, IUPs and failing PULs (spontaneous miscarriages) among women who presented to an early pregnancy unit. In particular, their M4 model (Condous *et al*., 2007a) utilizes two hCG levels: one taken at presentation and another taken ~48 h later. This model uses polytomous logistic regression (Hosmer and Lemeshow, 2000) to predict for each woman the probability of each type of pregnancy (EP, IUP, failing PUL) based on three measurements: log hCG average (natural log of the mean of the two hCG levels), hCG ratio centered (hCG level at 48 h divided by that at 0 h, minus its average) and the quadratic effect of the hCG ratio centered. The resulting probability estimates are then multiplied by weighting factors to obtain weighted predicted probabilities. The most likely outcome, defined by the highest weighted predicted probability, for each of the three possibilities for each woman is considered the predicted outcome.

The UK population for analysis consisted of 431 women: 228 (53%) with a failing PUL, 177 (41%) with an IUP and 26 (6%) with an EP. To be consistent with the application of the M4 model (Condous *et al*., 2007a), we limited the US population to women whose diagnosis was not definitive at presentation, who had two subsequent hCG values, ~48 h apart, and who were ultimately definitively diagnosed as having an EP, IUP or spontaneous pregnancy. This population of 604 women, 351 (58%) spontaneous miscarriage, 157 (26%) IUP and 96 (16%) EP, was labeled the ‘US population’.

Because of differences in inclusion criteria, diagnostic criteria and terminology between the USA and the UK, we also conducted additional analyses using an adjusted population where the inclusion criteria and definition of outcome were defined to more closely match the population described by Condous *et al*. (2007a). First, we reclassified certain outcomes. Condous *et al*. (2007a) and Kirk *et al*. (2007b) defined IUP to include all women with intrauterine gestation, regardless of viability. Therefore, using our subcategories of spontaneous miscarriage, we reclassified 11 women with an empty sac (anembryonic gestation), missed miscarriage (intrauterine fetal demise) or incomplete miscarriage (or retained products of conception) to an outcome of IUP. However, these women were subsequently excluded when we limited this analysis to include only women who received a non-diagnostic ultrasound at presentation, censoring cases where there was a suspicion for an EP or IUP based on non-definitive ultrasound criteria (−156 women), or when we limited the population to women with an initial hCG level below 10 000 IU/L (−10 women). Furthermore, recognizing that not all women in the US population returned for a subsequent follow-up exactly 2 days after presentation, we broadened our criteria to include women who had hCG readings at 1 or 3 days (54 and 52 women, respectively) after their initial presentation. For those who had a reading after 1 or 3 days but not after 2 days, a ‘2-day hCG’ was interpolated by assuming a linear slope of hCG change over time. This calculated ‘2-day hCG’ was used in modeling. This population was labeled the ‘adjusted US population’ and consisted of 544 women: 302 (56%) spontaneous miscarriage/failing PUL, 138 (25%) IUP and 104 (19%) EP.

When applying the model, we centered the terms for the hCG ratio and the quadratic effect of the ratio by subtracting the mean hCG ratio calculated from our own data (1.248). However, we measured model performance using the regression coefficients and weighting factors as reported by Condous *et al*. (2007a).

Data analysis

In reporting descriptive statistics (Table I), differences between outcome groups for continuous variables were measured using the non-parametric Kruskal–Wallis test due to the lack of normality of these variables. Differences between categorical variables were measured using Pearson’s χ^2.

Polytomous logistic regression was used to refit (i.e. retrain) M4. Odds ratios (ORs) and 95% confidence interval (CI) estimates were obtained to compare the effect of each variable in the model between the US and the UK populations.

Model performance of the original M4 in the various populations was assessed by calculating the area under the ROC curve (AUC). CIs for these AUCs were obtained using a logit-transform method (Pepe, 2003; Qin and Hotilovac, 2008). Predicted outcomes for each patient were obtained via a three-step process as described in the Condous *et al.* (2007a) paper. Model performance was then measured by calculating the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each possible outcome.

Results

Descriptive statistics of the two populations are presented in Table I. There were differences in average hCG values and the estimated gestational age at presentation between US and UK patients.

In Table II, we compare the changes in model parameter estimates obtained by refitting the M4 model to the original and adjusted US populations. The ORs are presented for each of the three variables used in the prediction model. Generally, the effects of the variables were less strong in the US populations. In predicting EP versus IUP, the ORs for log hCG average and hCG ratio centered were <1 in all three populations. This reflects the fact that patients with EPs are more likely to have lower hCG values and less likely to have hCG values that are increasing over time (resulting in a higher hCG ratio) than patients with IUPs.

The ORs for the comparison for spontaneous miscarriage/failing PUL to IUP were similar in direction of association in the UK compared with the US populations. However, in this case, the ORs were much lower and more highly significant in the UK population than in the US populations.

Finally, the comparison of EP versus spontaneous miscarriage shows additional differences between the effectiveness of the variables in the

Table I Baseline characteristics of the cohorts.

Scenario	Variable	Spontaneous miscarriage/failing PUL	IUP	EP	Overall *P*-value between groups
UK cohort (*N* = 431)		*n* = 228 (53%)	*n* = 177 (41%)	*n* = 26 (6%)	
	hCG ratio	0.41 (0.35)	2.25 (0.53)	1.16 (0.27)	<0.0001
	hCG average	137 (423)	804 (842)	711 (1069)	<0.0001
	Gestational age (days)	44 (14)	29 (5)	39 (7)	<0.0001
US cohort (*N* = 604)		*n* = 351 58%	*n* = 157 26%	*n* = 96 16%	
	hCG ratio	0.42 (0.49)	2.58 (0.99)	1.15 (0.75)	<0.0001
	hCG average	378 (984)	870 (1740)	350 (563)	<0.0001
	Gestational age (days)	47.0 (21.0)	34.0 (11.0)	39.0 (11.0)	<0.0001
Adjusted population (*N* = 544)		*n* = 302 56%	*n* = 138 25%	*n* = 104 19%	
	hCG ratio	0.42 (0.47)	2.69 (1.01)	1.23 (0.67)	<0.0001
	hCG average	289 (588)	658 (1220)	345 (608)	<0.0001
	Gestational age (days)	45.0 (16.0)	34.0 (8.0)	39.0 (12.0)	<0.0001

Values represent median (IQR).

Table II ORs from multivariate logistic regression.

Scenario	Variable	OR (95% CI)		
		EP versus IUP	Spontaneous miscarriage/ Failing PUL versus IUP	EP versus spontaneous miscarriage/failing PUL
UK cohort	Log hCG average	0.67 (0.36–1.21)	0.14 (0.05–0.30)*	4.94 (2.49–11.8)*
	hCG ratio centered	0.01 (<0.01 to 0.05)*	<0.01 (<0.01 to <0.01)*	70.2 (15.9–501.6)*
	Quadratic effect of hCG ratio centered	2.52 (0.46–5.83)	10.9 (5.36–23.6)*	0.23 (0.06–0.46)*
U.S. Population	Log hCG average	0.60 (0.47–0.76)*	0.71 (0.56–0.91)*	0.84 (0.71–0.99)*
	hCG ratio centered	0.19 (0.08–0.41)*	0.018 (0.01–0.04)*	10.15 (6.25–16.49)*
	Quadratic effect of hCG ratio centered	0.46 (0.26–0.84)*	2.10 (1.71–2.58)*	0.22 (0.13–0.38)*
Adjusted U.S. Population	Log hCG average	0.64 (0.49–0.83)*	0.66 (0.50–0.87)*	0.96 (0.80–1.13)
	hCG ratio centered	0.29 (0.13–0.65)*	0.03 (0.01–0.06)*	10.51 (6.41–17.22)*
	Quadratic effect of hCG ratio centered	0.41 (0.22–0.76)*	1.80 (1.20–2.72)*	0.23 (0.14–0.38)*

*OR is statistically significant.

US versus the UK population. In all cases, the OR for the hCG ratio is well above 1 and statistically significant, indicating that higher ratios (increasing hCG over time) are much more likely to reflect an EP than a spontaneous miscarriage. In the US populations, this OR was just over 10; however, in the UK, the OR was about seven times higher, with the upper CI reaching above 500, implying a superior ability to predict EP versus spontaneous miscarriage based on 2-day hCG ratios. With regard to the hCG average, the ORs of the two populations went in opposite directions. With the original UK population, higher hCG averages were associated with higher probability of EP versus spontaneous miscarriage (OR = 4.94, 95% CI = 2.49–11.8). In the US population, the comparable OR was 0.84, with the upper end of the CI at 0.99. In the adjusted US population, this result was not significant, in spite of a relatively tight CI (OR = 0.96, 95% CI = 0.80–1.13). This shows that in the US populations studied, a higher average hCG value is not associated with a higher probability of EP versus spontaneous miscarriage, in contrast to the results found in the UK population.

Table III Area under the ROC curve for prediction of each diagnosis.

Scenario	Prediction	Applied model
UK cohort	Failing PUL	0.988 (0.973–0.994)
	IUP	0.981 (0.962–0.990)
	EP	0.904 (0.789–0.960)
US population	Spontaneous miscarriage	0.885 (0.854–0.910)
	IUP	0.919 (0.883–0.945)
	EP	0.807 (0.757–0.849)
Adjusted US population	Spontaneous miscarriage	0.898 (0.867–0.922)
	IUP	0.915 (0.875–0.943)
	EP	0.831 (0.783–0.869)

In Table III, we present the AUCs that result from direct application of the original M4 model formulae. In all three populations tested, the AUC for prediction of EPs is significantly lower than the AUCs for the other two outcomes. For all three outcomes, the model performs worse with either of the US populations studied compared with performance using the original data. In spite of this decrease in performance relative to the UK population, all AUCs were between 0.8 and 0.9.

To obtain the data in Table IV, we present model performance for an application of the original M4 model, i.e. using the regression coefficients and weighting factors derived from the original population, to predict outcomes for the US populations. The contrast of predicted versus actual outcomes again highlights the lower ability of this model to distinguish between outcomes, especially between EPs and spontaneous miscarriage, in the US populations. This is particularly evident in the EP sensitivity numbers, which are only 49 and 55%, respectively, in the unadjusted and adjusted US populations. EP also has the lowest sensitivity of the three outcomes in the original UK population, but was still relatively high at 81%. The sensitivity for spontaneous miscarriage and IUP, and the specificity for IUP and EP were all lower in the US populations than in the original UK population, but these differences were relatively minor. The specificity for spontaneous miscarriage was much lower in the US populations (83%) than in the original population (98.5%), due primarily to the high number of EPs that the model predicted to be spontaneous miscarriage. The two indices for which the model worked better in the US populations were the PPV for EP and the NPV for IUP.

Discussion

In this study, we have tried to validate the logistic regression model M4 on a population external to that in which it was developed. We have shown that the M4 model developed on UK data performs worse when directly applied to a population of US women. Despite AUCs between 0.80 and 0.92, the prediction strategy of weighting probabilities aimed at a high detection of EP through the weighting, resulted in a sensitivity for EP of ~50%. In contrast, the AUCs were between 0.90 and 0.99 for the UK women, with a sensitivity of EP of 81%.

Table IV Predicted and actual outcomes.

Scenario	Actual outcome	Predicted outcome				Performance index			
		Spontaneous miscarriage/failing PUL	IUP	EP	Total	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
UK cohort	Spontaneous miscarriage/failing PUL	200	2	26	228	87.7	98.5	98.5	87.7
	IUP	3	155	19	177	87.6	97.2	95.7	91.8
	EP	0	5	21	26	80.8	88.9	31.8	98.6
	Total	203	162	66	431				
US population	Spontaneous miscarriage/failing PUL	286	14	51	351	81.4	83.0	86.9	76.4
	IUP	12	132	13	157	84.1	92.8	80.5	94.3
	EP	31	18	47	96	49.0	87.4	42.3	90.0
	Total	329	164	111	604				
Adjusted US population	Spontaneous miscarriage/failing PUL	251	11	40	302	83.1	83.1	86.0	79.8
	IUP	11	113	14	138	81.9	93.1	80.1	93.8
	EP	30	17	57	104	54.8	87.7	51.4	89.1
	Total	292	141	111	544				

Although M4's predictors are able to discriminate between the three outcomes in the US cohort as evidenced by the AUCs, the strength of the effects clearly differs from the UK cohort. This has caused the sensitivity for EP to be poor in the US data. This result is too low for the model to be used in clinical practice in its present form. However, significant differences in definitions used to define a PUL in the UK and US centers contribute to limitations and interpretation of these findings. As a result of these differences, adjustments were made trying to make the data compatible. The predictive model's performance modestly improved when adjustments were made for differences in definition and clinical practice.

Differences in definition and practice include that in the UK, a PUL was defined as no evidence of a pregnancy either inside or outside the uterus based on ultrasound. In the USA, the follow-up consists of women at risk for an EP in whom a diagnosis was not definitive at presentation. Therefore, the population evaluated included women noted to have a possible early gestational sac when a yolk sac or embryo could not be visualized. Moreover, failing pregnancies were also classified differently in the two cohorts. Although we attempted to make the UK and US populations of women at risk for EP as comparable as possible, we still noted significant differences.

These data also show that the population of women evaluated in the USA and UK may be different, independent of differences in definitions. It appears that women in the USA present later in the natural history of an IUP, as they were noted to have a greater gestational age (34 versus 29 days) but with a similar median hCG level compared with women in the UK. Interestingly, women in the USA with a spontaneous miscarriage also present at a higher gestational age but have an average hCG value almost 3-fold higher than that of women in the UK. Women in the USA with an EP have, on average, an almost identical gestational age as those in the UK, but a lower average hCG level. Finally, there were significant differences in the prevalence of EPs (16 and 6%) in the US and UK centers, respectively. It is not known whether more EPs had been visualized at ultrasonography during the first visit in the UK and thus had not been included in the PUL study group or whether the incidence of EPs is higher in the US population. Notwithstanding these differences, the two populations in general have an average estimated gestational age and hCG values typical of a woman with a symptomatic first-trimester pregnancy in which a model to predict outcome would be of value.

Specific examination of the parameters of the model demonstrates how well predictors may distinguish among the outcomes. OR estimates in the UK and US populations generally are in the same direction. However, the ORs are often substantially stronger in the UK population. For example, when comparing and EP versus spontaneous miscarriage/failing PUL, this means a given degree of change in hCG values over 48 h is more likely to distinguish between EPs and failed gestation in the UK population than the same degree of change in the original US population. The decrease in the ability to distinguish EP and spontaneous miscarriage/failing PUL in the US populations compared with the UK population is reflected in the lower AUC and predictive value for both outcomes. This difference is important as the result could be the misclassification of women with a miscarriage as an EP and women with an EP as a miscarriage. Clinically, this can result in overtreatment of women with a miscarriage with surgery or methotrexate, with inherent treatment morbidity, as well as potential rupture of an EP in a woman predicted to have a spontaneously resolving gestation.

The prediction of an IUP (based on sensitivity and AUC) was also lower in the USA, although this difference was not as great as for the other outcomes. Consistent prediction of an IUP likely reflects the predictable rise of hCG in women with a viable gestation (Barnhart *et al*., 2004a).

An alternative explanation for the discrepancy in the predictive ability of the mathematical model in the USA and the UK is that diagnosis and follow-up of women at risk for EP is not simply based on serial hCG values. Differences may arise for dissimilarities in the clinical management of women at risk. In the UK, asymptomatic as well as symptomatic patients were seen in a dedicated early pregnancy assessment unit by staff with a special interest in early pregnancy complications. Thus, in the UK, a qualified gynecologist or specialist nurse may also take into consideration subtle physical or ultrasound findings in determining follow-up. In contrast, in the US population, most ultrasound examinations were performed by a radiologist assigned to coverage of an emergency department with referral of patients to a gynecologist after the ultrasound had been performed. These differences may alter the prevalence of outcomes and the acuity of risk of subjects followed with outpatient surveillance and thus potentially alter the model's predictive ability.

Finally, the diversity of the natural history of an untreated non-viable gestation and/or EP may affect the ability of a mathematical model to predict ultimate outcome. We have demonstrated that the decline in hCG values in women with a spontaneous resolution of an abnormal gestation has a large range, likely due to variable production of hCG (Barnhart *et al*., 2004b). We have also demonstrated that there is no way to characterize the curves generated by serial hCG values of women with an EP as some rise, some fall and some change direction (Silva *et al*., 2006). Moreover, it is known that the rise and fall in serial hCG in women with an EP can mimic that expected for an ongoing IUP or that of a resolving abnormal gestation (Barnhart, 2009). Thus, it is possible that more data than two hCG values 48 h apart are needed to predict the ultimate outcome of women at risk for an EP (Seeber *et al*., 2006) or that prediction is not possible based on serial hCG values in some patients.

This study illustrates important difficulties in interpreting data derived from different health-care settings. Both UK and US centers have published extensively on the subject of early pregnancy complications, yet it is clear that both were defining PULs differently and examining patients in very different clinical settings. These differences only became clear after face-to-face meetings. We believe that it is important to share this experience as it shows the difficulties that may arise when applying clinical algorithms from the literature without absolute clarity regarding definitions and the population treated. In order to optimize models, it is clear that it is necessary to agree on a common classification for PUL and to prospectively collect data in multicenter studies with similar inclusion criteria. In this way, any model developed will be more generally applicable.

In conclusion, prediction of the location of a gestation would be a valuable clinical aid. However, the M4 model has a lower ability to predict an EP in this specific US population compared with the one in which it was developed. For the moment, continued surveillance, beyond 48 h, for women at risk for EP is necessary to make a definitive diagnosis and start optimal treatment in order to limit morbidity and mortality of the disease.

Authors' roles

K.T.B.: study conception, data analysis and interpretation, drafting of manuscript. M.D.S.: study conception, statistical analysis, data analysis and interpretation, drafting of manuscript. D.A.: statistical analysis, data analysis and interpretation, drafting of manuscript. M.R.: study conception, data analysis and interpretation, drafting of manuscript. T.M.: study conception, data analysis and interpretation, drafting of manuscript. B.C.: data analysis and interpretation, drafting of manuscript. S.H.: data collection, drafting of manuscript. E.K.: data collection, drafting of manuscript. G.C.: data collection, drafting of manuscript. D.T.: data analysis and interpretation, drafting of manuscript. T.B.: data analysis and interpretation, drafting of manuscript.

Funding

Funded by NIH R01-HD036455 (K.T.B.); Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC); Research Foundation—Flanders (FWO): FWO G.0302.07 (SVM), G.0341.07 (Data fusion); IWT: TBM070706-IOTA3; Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007–2011). B.C. is a postdoctoral researcher from the Research Foundation—Flanders (FWO).

References

Barnhart KT. Ectopic pregnancy. *N Engl J Med* 2009;**361**:379–387.

Barnhart KT, Sammel MD, Rinaudo PF, Zhou L, Hummel AC, Guo W. Symptomatic patients with an early viable intrauterine pregnancy: hCG curves redefined. *Obstet Gynecol* 2004a;**104**:50–55.

Barnhart K, Sammel MD, Chung K, Zhou L, Hummel AC, Guo W. Decline of serum human chorionic gonadotropin and spontaneous complete abortion: defining the normal curve. *Obstet Gynecol* 2004b; **104**:975–981.

Bignardi T, Condous G, Alhamdan D, Kirk E, Van Calster B, Van Huffel S, Timmerman D, Bourne T. The hCG ratio can predict the ultimate viability of the intrauterine pregnancies of uncertain viability in the pregnancy of unknown location population. *Hum Reprod* 2008; **23**:1964–1967.

Condous G, Lu C, Van Huffel SV, Timmerman D, Bourne T. Human chorionic gonadotrophin and progesterone levels in pregnancies of unknown location. *Int J Gynaecol Obstet* 2004a;**86**:351–357.

Condous G, Okaro E, Khalid A, Timmerman D, Lu C, Zhou Y, Van Huffel SV, Bourne T. The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. *Hum Reprod* 2004b;**19**:1900–1910.

Condous G, Kirk E, Lu C, Van Calster B, Van Huffel S, Timmerman D, Bourne T. There is no role for uterine curettage in the contemporary diagnostic workup of women with a pregnancy of unknown location. *Hum Reprod* 2006;**21**:2706–2710.

Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, Bourne T. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound Obstet Gynecol* 2007a;**29**:680–687.

Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, Bourne T. Clinical information does not improve the performance of mathematical models in predicting the outcome of pregnancies of unknown location. *Fertil Steril* 2007b;**88**:572–580.

Gevaert O, De Smet F, Kirk E, Van Calster B, Bourne T, Van Huffel S, Moreau Y, Timmerman D, De Moor B, Condous G. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Hum Reprod* 2006;**21**:1824–1831.

Hosmer DW, Lemeshow S. *Applied Logistic Regression*, 2nd edn. New York, NY: Wiley, 2000.

Kirk E, Condous G, Haider Z, Lu C, Van Huffel S, Timmerman D, Bourne T. The practical application of a mathematical model to predict the outcome of pregnancies of unknown location. *Ultrasound Obstet Gynecol* 2006;**27**:311–315.

Kirk E, Papageorghiou AT, Condous G, Tan L, Bora S, Bourne T. The diagnostic effectiveness of an initial transvaginal scan in detecting ectopic pregnancy. *Hum Reprod* 2007a;**22**:2824–2828.

Kirk E, Condous G, Van Calster B, Van Huffel S, Timmerman D, Bourne T. Rationalizing the follow-up of pregnancies of unknown location. *Hum Reprod* 2007b;**22**:1744–1750.

Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press, 2003.

Qin G, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat Methods Med Res* 2008;**17**:207–221.

Seeber BE, Barnhart KT. Suspected ectopic pregnancy. *Obstet Gynecol* 2006;**107**:399–413.

Seeber BE, Sammel MD, Guo W, Zhou L, Hummel A, Barnhart KT. Application of redefined human chorionic gonadotropin curves for the diagnosis of women at risk for ectopic pregnancy. *Fertil Steril* 2006; **86**:454–459.

Silva C, Sammel MD, Zhou L, Gracia C, Guo W, Hummel AC, Barnhart K. Human chorionic gonadotropin profile for women with an ectopic pregnancy. *Obstet Gynecol* 2006;**107**:605–610.

Van Calster B, Condous G, Kirk E, Bourne T, Timmerman D, Van Huffel S. An application of methods for the probabilistic three-class classification of pregnancies of unknown location. *Artif Intell Med* 2009;**46**: 139–154.