Hum. Reprod. Advance Access published May 10, 2016

Human Reproduction, Vol.0, No.0 pp. 1–11, 2016

doi:10.1093/humrep/dew105

human reproduction

ORIGINAL ARTICLE *Early pregnancy*

The clinical performance of the M4 decision support model to triage women with a pregnancy of unknown location as at low or high risk of complications

S. Bobdiwala[1,†], S. Guha[1,2,†], B. Van Calster[3], F. Ayim[4], N. Mitchell-Jones[5], M. Al-Memar[1], H. Mitchell[4], C. Stalder[1], C. Bottomley[5], A. Kothari[4], D. Timmerman[3,6], and T. Bourne[1,3,6,*]

[1]Tommy's National Early Miscarriage Research Centre, Queen Charlottes & Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, UK [2]West Middlesex University Hospital, Twickenham Road, Isleworth, London TW7 6AF, UK [3]Department of Development and Regeneration, KU Leuven, Herestraat 49 Box 7003, Leuven B-3000, Belgium [4]Hillingdon Hospital, Pield Heath Road, Uxbridge UB8 3NN, UK [5]Chelsea & Westminster Hospital, 329 Fulham Road, London SW10 9NH, UK [6]Department of Obstetrics and Gynaecology, University Hospitals Leuven, Campus Gasthuisberg, KU Leuven, Belgium

*Correspondence address. E-mail: t.bourne@imperial.ac.uk

Submitted on November 5, 2015; resubmitted on March 24, 2016; accepted on April 7, 2016

STUDY QUESTION: What are the adverse outcomes associated with using the M4 model in everyday clinical practice for women with pregnancy of unknown location (PUL)?

SUMMARY ANSWER: There were 17/835 (2.0%) adverse events and no serious adverse events associated with the performance of the M4 model in clinical practice.

WHAT IS KNOWN ALREADY: The M4 model has previously been shown to stratify women classified as a PUL as at low or high risk of complications with a good level of test performance. The triage performance of the M4 model is better than single measurements of serum progesterone or the hCG ratio (serum hCG at 48 h/hCG at presentation).

STUDY DESIGN, SIZE, DURATION: A prospective multi-centre cohort study of 1022 women with a PUL carried out between August 2012 and December 2013 across 2 university teaching hospitals and 1 district general hospital.

PARTICIPANTS/MATERIALS, SETTING, METHODS: All women presenting with a PUL to the early pregnancy units of the three hospitals were recruited. The final outcome for PUL was either a failed PUL (FPUL), intrauterine pregnancy (IUP) or ectopic pregnancy (EP) (including persistent PUL (PPUL)), with EP and PPUL considered high-risk PUL. Their hCG results at 0 and 48 h were entered into the M4 model algorithm. If the risk of EP was $\geq$5%, the PUL was predicted to be high-risk and the participant was asked to re-attend 48 h later for a repeat hCG and transvaginal ultrasound scan by a senior clinician. If the PUL was classified as 'low risk, likely failed PUL', the participant was asked to perform a urinary pregnancy test 2 weeks later. If the PUL was classified as 'low risk, likely intrauterine', the participant was scheduled for a repeat scan in 1 week. Deviations from the management protocol were recorded as either an 'unscheduled visit (participant reason)', 'unscheduled visit (clinician reason)' or 'differences in timing (blood test/ultrasound)'. Adverse events were assessed using definitions outlined in the UK Good Clinical Practice Guidelines' document.

MAIN RESULTS AND THE ROLE OF CHANCE: A total of 835 (82%) women classified as a PUL were managed according to the M4 model (9 met the exclusion criteria, 69 were lost to follow-up, 109 had no hCG result at 48 h). Of these, 443 (53%) had a final outcome of FPUL, 298 (36%) an IUP and 94 (11%) an EP. The M4 model predicted 70% (585/835) PUL as low risk, of which 568 (97%) were confirmed as FPUL or IUP. Of the 17 EP and PPUL misclassified as low risk, 5 had expectant management, 7 medical management with methotrexate and 5 surgical intervention.

[†] To be considered as joint first authors.

Nineteen PUL had an unscheduled visit (participant reason), 38 PUL had an unscheduled visit (clinician reason) and 68 PUL had deviations from protocol due to a difference in timing (blood test/ultrasound).

Adverse events were reported in 26 PUL and 1 participant had a serious adverse event. A total of 17/26 (65%) adverse events were misclassifications of a high risk PUL as low risk by the M4 model, while 5/26 (19%) adverse events were related to incorrect clinical decisions. Four of the 26 adverse events (15%) were secondary to unscheduled admissions for pain/bleeding. The serious adverse event was due to an incorrect clinical decision.

LIMITATIONS, REASONS FOR CAUTION: A limitation of the study was that 69/1022 (7%) of PUL were lost to follow-up. A 48 h hCG level was missing for 109/1022 (11%) participants.

WIDER IMPLICATIONS OF THE FINDINGS: The low number of adverse events (2.0%) suggests that expectant management of PUL using the M4 prediction model is safe. The model is an effective way of triaging women with a PUL as being at high- and low-risk of complications and rationalizing follow-up. The multi-centre design of the study is more likely to make the performance of the M4 model generalizable in other populations.

STUDY FUNDING/COMPETING INTEREST(S): None.

TRIAL REGISTRATION NUMBER: Not applicable.

Key words: ectopic pregnancy / miscarriage / pregnancy of unknown location / decision support techniques / ultrasonography / triage / adverse events

Introduction

Pregnancy of unknown location (PUL) is a common management problem in early pregnancy (Bottomley et al., 2009). It is a clinical scenario defined as when a urinary pregnancy test is positive but neither an intrauterine pregnancy (IUP) nor extra-uterine pregnancy can be visualized on a transvaginal ultrasound scan (TVS). The incidence of PUL depends to an extent on the quality of ultrasound in a unit and has been reported to be between 8 and 31% (Hahlin et al., 1995; Banerjee et al., 1999; Kirk et al., 2007, 2009, 2014; van Mello et al., 2012). Due to a lack of standardized protocols, the management of PUL varies and may be stressful for women who are often subjected to repeated blood tests and scans before the final outcome of the pregnancy is known. In general, research in this field has focused on identifying a test that enables clinicians to identify ectopic pregnancies (EP) within the overall population of PUL.

Clinically, the most common method used to predict PUL outcome is the hCG ratio (serum hCG at 48 h/hCG 0 h). An hCG ratio of <0.87 is associated with a failing PUL and >1.66 with a likely viable IUP. A suboptimal rise between 0.87 and 1.66 indicates an increased risk of EP (Condous et al., 2004a). Prediction models have also been developed to aid the management of women classified as having a PUL, including logistic regression models and Bayesian networks. These have considered biochemical variables such as serum hCG, the hCG ratio and progesterone levels, in combination with ultrasound-based variables such as the endometrial thickness and clinical parameters such as vaginal bleeding (Banerjee et al., 2001; Condous et al., 2004b, 2007a,b; Gevaert et al., 2006). The most widely evaluated prediction model developed is M4 (Condous et al., 2007c; Van Calster et al., 2013), which is based on the initial serum hCG and the hCG ratio.

Currently, the focus for management of PUL has shifted from identifying the location of a pregnancy to classifying them as being at low or high risk of complications. Low-risk PUL are failing pregnancies and early viable IUP, with EP and persistent PUL (PPUL) constituting the high-risk group. Using this approach, follow-up can be reduced to a minimum for low-risk PUL and resources focused on PUL classified at high risk of EP.

The M4 logistic regression model uses both the initial serum hCG and the hCG ratio to assign women with a PUL into ‘low’ or ‘high’ risk groups. Use of this model has the potential to reduce follow up in 70% of PUL with a negative predictive value (NPV) of 97.5% (Van Calster et al., 2013). In a further study on 1271 PUL, Guha et al. (2014) compared the performance of serum progesterone, the hCG ratio and the M4 model. The M4 model performed better than both the hCG ratio and progesterone alone. The M4 model correctly classified 84% of EP as high risk with the highest odds ratio of all three tests.

Whilst the diagnostic performance of the M4 model has been validated both temporally and externally, a prospective assessment of any complications associated with its use in ‘real world’ clinical practice has not been carried out. In this study, the primary aim was to assess the number of complications or adverse outcomes associated with using the M4 model in everyday practice for triaging PUL as either low or high risk. The secondary aim was to further prospectively validate the diagnostic performance of the M4 model on a new population of women with a PUL.

Materials and Methods

Design and settings

This was a prospective multi-centre cohort study on consecutive women classified as having a PUL on presentation to the Early Pregnancy Assessment Units (EPAU) of three London, UK, hospitals. Two were university teaching hospitals: Queen Charlottes' and Chelsea Hospital (QCCH) and Chelsea and Westminster (C&W). The third was an outer London district general hospital: Hillingdon Hospital (HH). Data were collected between August 2012 and December 2013 at QCCH (16 months); February to September 2013 at HH (7 months) and March to September 2013 at C&W (6 months). Following previous publications (Van Calster et al., 2013; Guha et al., 2014), use of the M4 model had been incorporated into the routine clinical protocols in all three centres. Accordingly, after discussion with the research and development departments within the respective National Health Service hospitals, the study was registered as an audit, having been advised that ethical approval was not required as there was no change to the routine clinical management. The following specific inclusion and exclusion criteria were set.

Inclusion criteria—any woman seen that was classified as a PUL at the initial visit to the EPAU and was haemodynamically stable (and therefore fit for outpatient management).

Exclusion criteria—any woman seen that was not classified as a PUL at the initial visit to the EPAU (e.g. EP); any participant unsuitable for outpatient management with serial serum hCG levels (e.g. haemodynamically unstable); final diagnosis of a molar pregnancy (as it is known that serum hCG levels in this cohort of women behave in a way that will not be interpretable to any mode of management risk-stratifying PUL).

At presentation, all women were seen and a TVS performed by a trained sonographer (nurse specialist, sonographer or gynaecologist). A history was taken which included past history of EP and their method of treatment, amount of vaginal bleeding (expressed as a pictorial bleeding assessment score of 1–4) (Bottomley *et al.*, 2009) and pain (expressed using a 10-cm visual analogue scale).

Women were classified as having a PUL following a TVS according to the definition of PUL contained within a recent consensus paper (Barnhart *et al.*, 2011).

Serum hCG measurement

All patients had blood samples taken for serum hCG at the initial visit and 48 h later. At QCCH and C&W, blood samples were assayed using the Abbott hCG assay run onboard the Abbott Architect i2000SR instrument (Abbott Laboratories, Abbott Park, IL, USA). This is a two-step, non-competitive, two-site type immunoassay that employs two mouse monoclonal antibodies and CMIA 'Chemiflex^{TM} technology'. At HH, the UniCel DxI 800 Immunoassay System was used. It also uses chemiluminescent technology.

Triage using the M4 model

The M4 prediction model algorithm was embedded into a simple protected excel file, and all staff, including junior doctors on call in the evening and at weekends, had access to this file. On being classified as a PUL, women had the initial serum hCG level measured, and the results entered onto the excel spreadsheet. They were then asked to return 48 h later for a repeat serum hCG. Participants were also asked to telephone or return to the unit in the event of any deterioration in their clinical condition, such as increased levels of pain or vaginal bleeding. On entering the second serum hCG value, the model estimated the risk that the PUL was a failed PUL (FPUL), IUP or EP. The PUL was classified as 'high risk' (probable EP) if the predicted risk of EP was ≥5% (Van Calster *et al.*, 2013). When the risk was <5%, the PUL was classified as 'low risk, probable FPUL' or low risk, probable IUP'. These classifications were linked to agreed management plans. Women with a 'low risk probable FPUL' were advised to perform a urine pregnancy test in 2 weeks with telephone follow-up by the unit to confirm a negative result. If the urine pregnancy test was positive after 2 weeks, the women were brought back for a repeat serum hCG and TVS if required. The 'low risk probable IUP' cases were followed up with a repeat TVS after 1 week in order to confirm an IUP. Women classified as 'high risk probable EP' were advised to return within 48 h for a repeat ultrasound scan by a senior clinician (Fig. 1). If they remained a PUL, a repeat serum hCG was measured and the woman was managed according to the clinical scenario. Clinicians were told that the model was a guide to management and that it should not be followed blindly if they felt the clinical situation demanded a different management plan. The M4 model gave guidance on follow up but not on whether there was a need for intervention. The decision for intervention was always made by the responsible clinical team.

Reference standard

The final outcome was the outcome of the pregnancy at the 11–14 week dating scan. The outcomes were identified as: (i) viable IUP with visible fetal heart activity; (ii) non-viable IUP (where an IUP was identified on ultrasound but miscarried by the time of the dating scan); (iii) failing pregnancy (women who had a negative pregnancy test on follow up after 2 weeks); (iv) EP, defined as an extra-uterine mass visualized using ultrasonography. In decreasing order of specificity, the appearance could be a tubal ring with a yolk sac and embryo, a tubal ring with a yolk sac only, a tubal ring without central identifying features, an inhomogeneous or non-cystic adnexal mass sometimes known as the 'blob' sign (Condous *et al.*, 2005; Levine, 2007; Kirk *et al.*, 2008); (v) persistent PUL (women where the ultrasound diagnosis remained as a PUL over the course of 2 weeks with at least three hCG levels that did not change by more than 15%).

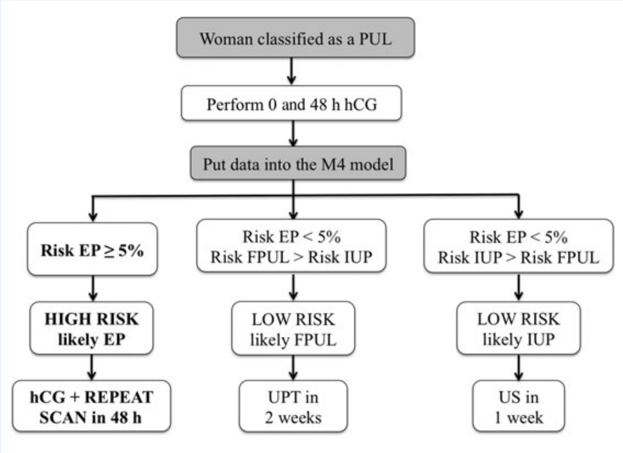

Figure 1 The three possible predictions and management plans as recommended by the M4 model after inputting the initial (0 h) and 48 h serum hCG results. PUL, pregnancy of unknown location; EP, ectopic pregnancy; FPUL, failed pregnancy of unknown location; IUP, intrauterine pregnancy; US, ultrasound; UPT, urine pregnancy test.

Deviations from protocol and adverse events

Protocol deviations: data were collected on any deviations from the management recommended by the M4 model and the reasons for those deviations. Data were also collected for interventions that took place (e.g. surgical management of miscarriage or laparoscopy) and whether these were planned or unplanned. The total number of blood tests and scans required prior to making a final diagnosis was recorded. Protocol deviations were subdivided into either 'minor' or 'major'.

A minor protocol deviation was described for three reasons:

First, an unscheduled visit (participant reason) where women made a decision to attend the EPAU or accident and emergency departments due to pain or vaginal bleeding rather than adhere to their follow-up plan.

Second, an unscheduled visit (clinician reason) where clinicians made a decision to alter the follow-up plan. This may have been based on an assessment of the overall clinical situation, their own interpretation of the serum hCG results or of the ultrasound images from the participant.

The third reason was incorrect timing (blood test/ultrasound), where the timing of a follow-up blood test or TVS was altered due to a participant/clinician decision or the presence of a weekend/bank holiday (all three EPAUs are only open Monday–Friday, 09:00–17:00).

A major protocol deviation was classified as a participant who met the exclusion criteria for using the M4 model and should not have been included in the study. This included women where a diagnosis of an EP had been made from the outset or the participant had a molar pregnancy confirmed at a later date (see Fig. 2 for a summary).

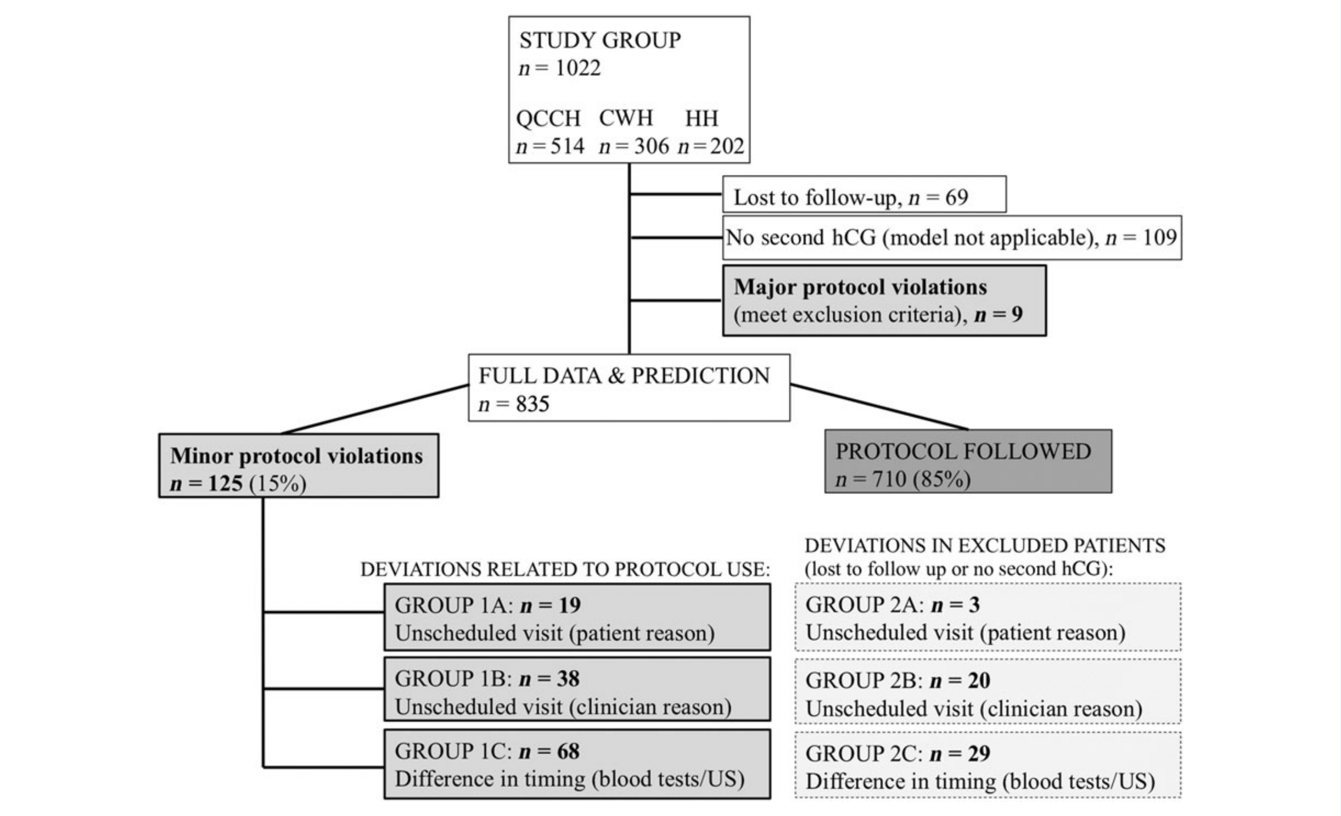
STUDY GROUP
n = 1022
QCCH CWH HH
n = 514 n = 306 n = 202
Lost to follow-up, n = 69
No second hCG (model not applicable), n = 109
Major protocol violations
(meet exclusion criteria), n = 9
FULL DATA & PREDICTION
n = 835
Minor protocol violations
n = 125 (15%)
PROTOCOL FOLLOWED
n = 710 (85%)
DEVIATIONS RELATED TO PROTOCOL USE:
GROUP 1A: n = 19
Unscheduled visit (patient reason)
GROUP 1B: n = 38
Unscheduled visit (clinician reason)
GROUP 1C: n = 68
Difference in timing (blood tests/US)
DEVIATIONS IN EXCLUDED PATIENTS
(lost to follow up or no second hCG):
GROUP 2A: n = 3
Unscheduled visit (patient reason)
GROUP 2B: n = 20
Unscheduled visit (clinician reason)
GROUP 2C: n = 29
Difference in timing (blood tests/US)

Figure 2 Flowchart of patient recruitment and protocol deviations.

Adverse events: we used the definitions and criteria for adverse events contained in the national Good Clinical Practice (GCP) guidelines (Medicines and Healthcare products Regulatory Agency (MHRA), 2014). These cover the legislation, guidance and good practice relating to the conduct of clinical trials in the UK. They define an 'adverse event' (AE) as 'any untoward medical occurrence in a patient or clinical study subject'. A 'serious adverse event' (SAE) is defined as 'any untoward and unexpected medical occurrence or effect that results in death (or) is life-threatening'. We extended the definition of a SAE to include any untoward occurrence that may have resulted in the death of a potentially viable pregnancy.

Statistical analysis

Statistical analysis was performed with SAS v9.4 (SAS Institute, Cary, NC, USA). The performance of the M4 decision support model was assessed by a cross-tabulation of reference standard (EP/PPUL versus FPUL/IUP) and classification (high risk versus low risk). The aim was to classify a large group of PUL as low risk whilst at the same time predicting most EP as high risk. Hence the main evaluation measures were the overall percentage of PUL classified as low risk, sensitivity (percentage of EP/PPUL classified as high risk) and NPV (percentage of non-EP among PUL classified as low risk). In addition, the false positive rate was computed (percentage of non-EP classified as high risk, this is 1 minus specificity). 95% confidence intervals (CI) were calculated using Wilson's score method (Newcombe, 1998a).

Performance was compared with that obtained for other triage protocols: a single visit protocol using a serum progesterone level of ≤10 nmol/l to assume low risk (Cordina et al., 2011), and hCG ratio cut-offs of <0.87 or >1.66 to assume low risk (Condous et al., 2006; Bignardi et al., 2008). Results for the M4 approach versus the progesterone or hCG ratio triage tools were compared using 95% CIs on the difference of overall percentage of PUL classified as low risk, sensitivity and false positive rate. The CIs were obtained using method 10 from Newcombe (1998b).

Discrimination and calibration performance of the predicted risks given by M4 were evaluated. Discrimination was assessed with the Polytomous Discrimination Index (PDI), which is an extension of the area under the receiver operating characteristic curve (AUC) for outcomes with more than two categories (Van Calster et al., 2012), and with the AUC for EP versus FPUL/IUP. Calibration of the predicted risk of EP was assessed using a flexible calibration analysis for outcomes with more than two categories (Van Hoorde et al., 2014). The analysis was first carried out for the complete cases, and then for all patients with single stochastic imputation of missing data. More information about the imputation method is given in Supplementary data.

Results

A total of 1022 women were classified as a PUL during the study period, 514 from QCCH, 306 from C&W and 202 from HH. Nine participants met the exclusion criteria and were omitted from the final analysis, while 69 (7%) participants were lost to follow up: 39 (8%) from QCCH, 26 (9%) from C&W and 4 (2%) from HH. The 48-h serum hCG was missing for 109 of the remaining 944 (11%) participants: 54/469 (12%) from QCCH; 42/279 (15%) from C&W and 13/196 (7%) from HH. The main reason for missing 48-h hCG results was a low (<50 IU/l) initial hCG level (68 participants: 28 QCCH, 33 C&W, 7 HH). In these cases, clinicians deviated from the protocol by asking women to carry out a urinary pregnancy test after 2 weeks rather than a serum hCG at 48 h. In all cases, the urinary pregnancy test was negative. Further reasons for missing 48-h hCG results were: failure to attend for the 48-h repeat blood test ($n = 28$), visualization of either an intra- or extrauterine gestation sac on a repeat scan ($n = 9$; 3 of these were an EP and 6 an IUP), a need for intervention prior to the second blood test ($n = 2$; one participant was admitted with heavy vaginal bleeding without haemodynamic compromise but required surgical management of a miscarriage and had a final outcome of a non-viable IUP; the other participant presented with pain and had a laparoscopy which confirmed an unruptured EP), suspected pseudosac at the initial scan so underwent ultrasonography 48 h later showing an intrauterine gestation sac with a yolk sac rather than repeating the hCG blood test—the final outcome was a non-viable IUP ($n = 1$), and loss of blood samples ($n = 1$).

A total of 835/1022 (82%) PUL were included in the final analysis (Fig. 3). In 443 (53%) the final outcome was a FPUL, there were 298 (36%) IUP and 94 EP (11%). Table I summarizes symptoms, initial serum hCG and hCG ratios stratified by final outcome.

Interventions and safety: EPs misclassified by the PUL model

There were a total of 94 EP in the study. Amongst the 17 EP misclassified as low risk (including 2 PPUL), 5 (29%) were managed expectantly, 7 (41%) had methotrexate and 5 (29%) had a laparoscopic salpingectomy (Table II). Similar management was seen in the 77 EP correctly predicted as high risk, i.e. 19 (25%) underwent expectant management, 28 (36%) medical management with methotrexate and 23 (30%) surgical management with laparoscopy (Table III).

Protocol deviations

In 9 of the total cohort of 1022 cases there was a 'major protocol deviation'. The recommended management according to the risk assessment given by the M4 model was followed in 710 out of the 835 participants included in the analysis (85%). In 125 (16%) participants there was a 'minor protocol deviation'. Nineteen participants had an unscheduled visit because of a participant decision, 38 secondary to a clinical decision by a doctor and 68 due to issues with the timing of either follow-up blood tests or ultrasound scans (see Fig. 2).

Of the 19 participants that had an unscheduled attendance, 1 had a final outcome of a viable IUP, 5 were a non-viable IUP, 5 were a FPUL and 8 had a final outcome of an EP/PPUL. Six of these eight EP/PPUL had been incorrectly classified as low risk by the M4 model. Of these six cases, three underwent laparoscopy, two were treated successfully with methotrexate and one had persistent hCG levels despite treatment with methotrexate and underwent a laparoscopic salpingectomy as a second line treatment. The three cases that had a laparoscopy as first line treatment re-presented with pain but none of the cases were found at surgery to have a ruptured EP. Two of 19 women had a negative laparoscopy having re-presented with pain and had a final outcome of a viable IUP and FPUL. Four of 19 cases re-presented with heavy vaginal bleeding and required emergency surgical management of miscarriage (IUP confirmed on histology).

Clinical safety data

Following the definitions of the GCP guidelines (MHRA, 2014), 26 women had AEs and 1 woman had a SAE (see Fig. 4 for a summary).

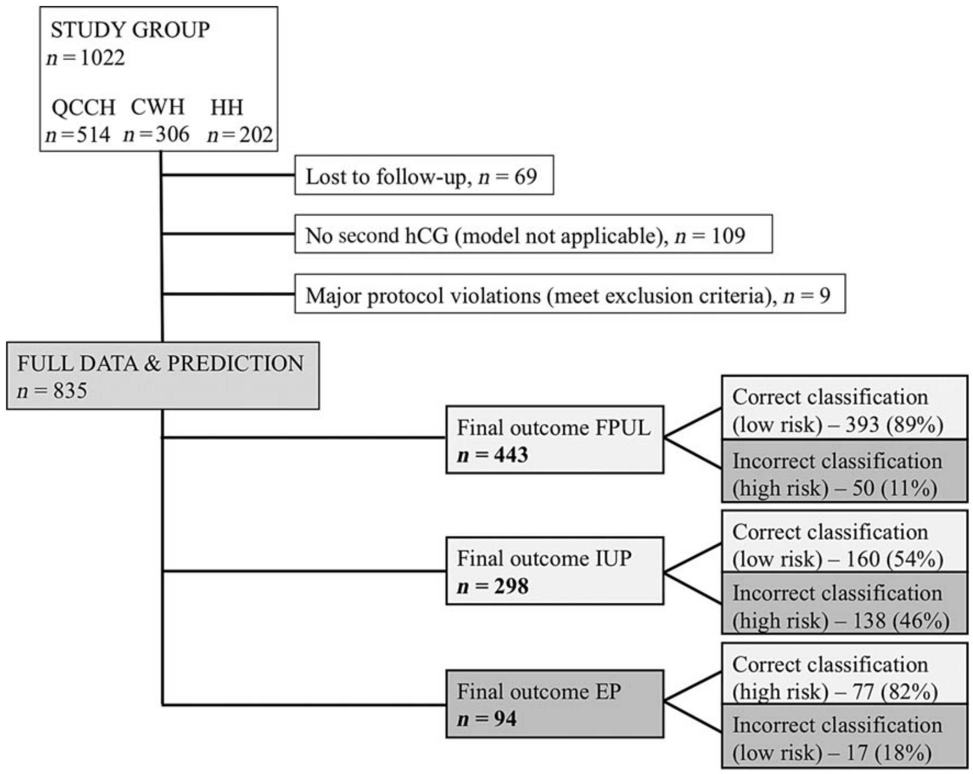

Figure 3 Flowchart of final outcome data and correct versus incorrect risk stratification according to the M4 model.

Table I **Data on demographics, symptoms and serum biochemistry results for patients in the study with pregnancy of unknown location (PUL).**

Characteristic	All (*n* = 835)	Failing PUL (*n* = 443)	IUP (*n* = 298)	EP (*n* = 94)
Age (years)	32 (28–36)	33 (28–38)	31 (26–35)	33 (29–36)
Initial serum hCG (IU/l)	571 (183–1914)	418 (115–1937)	825 (363–2081)	486 (189–1187)
48 h serum hCG (IU/l)	462 (123–1707)	155 (48–510)	1680 (701–3439)	504 (251–1516)
hCG ratio	0.74 (0.33–1.84)	0.35 (0.24–0.46)	2.04 (1.55–2.42)	1.18 (0.93–1.49)
Initial progesterone (nmol/l)	12 (4–40)	4 (2–9)	49 (31–64)	19 (7–32)
Mean pain score	2.1	2.2	2.0	2.1
Vaginal bleeding				
No bleeding (0)	211 (25%)	22 (5%)	169 (57%)	20 (21%)
Minimal bleeding (1)	230 (28%)	114 (26%)	75 (25%)	41 (44%)
Moderate bleeding (2)	187 (22%)	130 (29%)	33 (11%)	24 (26%)
Soaked sanitary towel (3)	92 (11%)	80 (18%)	7 (2%)	5 (5%)
Clots or flooding (4)	115 (14%)	97 (22%)	14 (5%)	4 (4%)
Mean score	1.6	2.3	0.7	1.3
History of EP				
Any	45 (5%)	10 (2%)	25 (8%)	10 (11%)
Laparoscopy	32 (4%)	7 (2%)	19 (6%)	6 (6%)
Methotrexate	7 (1%)	2 (<1%)	3 (1%)	2 (2%)
Expectant	6 (1%)	1 (<1%)	3 (1%)	2 (2%)
Indication for scan				
Bleeding and pain	432 (52%)	299 (67%)	79 (27%)	54 (57%)
Bleeding only (no pain)	184 (22%)	119 (27%)	47 (16%)	18 (19%)
Pain only (no bleeding)	159 (19%)	13 (3%)	131 (44%)	15 (16%)
Previous ectopic	14 (2%)	4 (1%)	9 (3%)	1 (1%)
Unsure dates	10 (1%)	2 (<1%)	8 (3%)	0 (0%)
Previous miscarriage	2 (<1%)	0 (0%)	2 (1%)	0 (0%)
Maternal reassurance	1 (<1%)	0 (0%)	0 (0%)	1 (1%)
Other	33 (4%)	6 (1%)	22 (7%)	5 (5%)

Results are shown as median (interquartile range) for age and biomarkers, mean score for pain and vaginal bleeding, and *n* (%) for vaginal bleeding, history of ectopic pregnancy (EP) and indication for scan. There were 248 missing values for initial progesterone (30%), 177 for pain score (21%) and 2 for history of EP (0.2%).
IUP, intrauterine pregnancy.

AEs: M4 model-related

These included the 17 women with an EP or PPUL who were incorrectly predicted as low risk (see Table II).

AEs: clinician-related

Five participants had negative laparoscopies secondary to a clinical decision by a doctor, i.e. an adnexal mass or blood was seen in the pelvis seen on an ultrasound scan (none of these participants had pelvic pain). Four of these participants were classified as low risk by the model and ultimately had a final outcome of viable IUP ($n = 1$), non-viable IUP ($n = 1$) and FPUL ($n = 2$). For one participant who was classified as high risk by the model, eventually a non-viable IUP was confirmed on histology after undergoing surgical management of miscarriage.

AEs: participant-related

Two participants had an unscheduled admission with heavy vaginal bleeding but were managed expectantly without the need for emergency surgery. They had a final outcome of FPUL and a non-viable IUP. A further two women had an unscheduled attendance to the hospital for pain and were appropriately admitted for laparoscopies but both were negative.

SAEs

One participant had a SAE. She had serum hCG measurements performed 24 rather than 48 h apart and the M4 model was incorrectly used. An invalid risk prediction of 'high risk' was given and the woman was incorrectly given methotrexate for a pregnancy that was later found to be an IUP and was terminated due to the risks of methotrexate on a developing fetus.

Number of blood tests and scans

Overall the mean number of scans per woman was 2.1, and blood tests 2.6. A breakdown of the number of scans and blood tests according to the final PUL outcome is shown in Table IV.

Table II The 17 high risk patients (15 ectopic pregnancies and 2 PPUL) which were misclassified as 'low risk' by the model and the interventions they had.

Patient	hCG 0 h	hCG 48 h	hCG ratio	Prog 0 h	M4 model prediction of outcome	Actual final outcome	Intervention	Indication for intervention
1	626	408	0.652		FAILED PUL	Ectopic pregnancy	Expectant management	
2	330	700	2.121	19	IUP	Ectopic pregnancy	Expectant management	
3	485	281	0.579	44	FAILED PUL	Ectopic pregnancy	Expectant management	
4	1550	436	0.281	54	FAILED PUL	Ectopic pregnancy	Expectant management	
5	1805	926	0.513	16	FAILED PUL	Ectopic pregnancy	Expectant management	
6	84	256	3.048	28	IUP	Ectopic pregnancy	Methotrexate	Ectopic seen on 1 week follow-up ultrasound
7	105	83	0.790	4	FAILED PUL	Ectopic pregnancy	Methotrexate	Initial expectant management but hCG did not decline sufficiently
8	129	340	2.636	7	IUP	Ectopic pregnancy	Methotrexate	Later rise in hCG and ectopic seen on ultrasound
9	76	184	2.421	76	IUP	Ectopic pregnancy	Methotrexate	PPUL on 1 week follow-up ultrasound
10	525	208	0.396	18	FAILED PUL	Ectopic pregnancy	Methotrexate	Initial expectant management but hCG did not decline sufficiently
11	187	132	0.706		FAILED PUL	Persisting PUL	Methotrexate	Negative laparoscopy and SMM, methotrexate given for persistent hCG
12	182	126	0.692	20	FAILED PUL	Persisting PUL	Methotrexate	Two methotrexate doses required
13	1164	2496	2.144	71	IUP	Ectopic pregnancy	Laparoscopy	Surgery as methotrexate not suitable (renal disease)
14	438	941	2.148	26	IUP	Ectopic pregnancy	Laparoscopy	Patient chose to have surgery
15	235	624	2.655		IUP	Ectopic pregnancy	Laparoscopy	Re-admitted with pain
16	162	389	2.401		IUP	Ectopic pregnancy	Laparoscopy	Re-admitted with pain
17	64	40	0.625	37	FAILED PUL	Ectopic pregnancy	Laparoscopy	Patient chose to have surgery

Prog, progesterone; PPUL, persistent pregnancy of unknown location; SMM, surgical management of miscarriage.

Evaluation of M4 model performance

The M4 model predicted 585/835 (70%) women with a PUL as low risk and was correct in 568/585 (97%) of cases. A total of 77 of 94 (82%) PUL with a final outcome of EP were correctly classified as high risk, and 173 of 741 (23%) FPUL or IUP were also classified as high risk. We found that 393 FPULs (89%) were correctly classified as 'low risk, probable FPUL, and finally 160 IUPs (54%) were appropriately classified as 'low risk, probable IUP'.

The model predicted 250 women with a PUL to be high risk. Of this cohort, 77 (31%) were correctly predicted as high risk and had an EP, 124 (50%) had a final diagnosis of an IUP and 49 (20%) a final diagnosis of a FPUL. Seventeen (18%) of the high risk EP/PPUL were misclassified as low risk (see Table II) and 173 (24%) of the IUP/FPUL were misclassified as high risk.

The performance of the M4 model and how it compares with triage protocols based on serum progesterone and hCG ratio is shown in Table V. The progesterone protocol was only applied to participants from QCCH and HH because progesterone was measured consistently in these centres (in 560 of 598, 94%). When hCG ratio cut-offs were used, sensitivities for EP and the false positive rate were lower. When a progesterone level of >10 nmol/l was used to classify participants as high risk, fewer PUL were classified as low risk and this prediction was more likely to be incorrect. More importantly, a greater number of EP were classified as low risk, whilst more FPUL and IUP were put in the high risk category.

The AUC to predict EP versus FPUL/IUP using M4 was 0.84 (0.79–0.87). The PDI, the AUC extension for multicategory outcomes, was 0.80. The calibration analysis suggested that the risk of EP was clearly

Table III An outline of all the interventions performed on patients included in the study. The risk prediction given by the M4 model, the final outcome at 12 weeks gestation and the specific intervention performed are given.

Outcome	Intervention	M4 triage		
		High-risk cases	Low-risk cases	Total number of cases
EP	SMM	6 (100%)	0	6
	Expectant	19 (79%)	5	24
	MTX	28 (80%)	7	35
	Laparoscopy	23 (82%)	5	28
	DNA	1	0	1
FPUL/IUP	None/expectant	148 (21%)	549	697
	TOP on patient request	3 (27%)	8	11
	SMM	19 (76%)	6	25
	MTX	1 (100%)	0	1
	Laparoscopy	1 (25%)	3	4
	DNA	1	1	2

FPUL, failed pregnancy of unknown location; MTX, methotrexate; TOP, termination of pregnancy; DNA, did not attend.

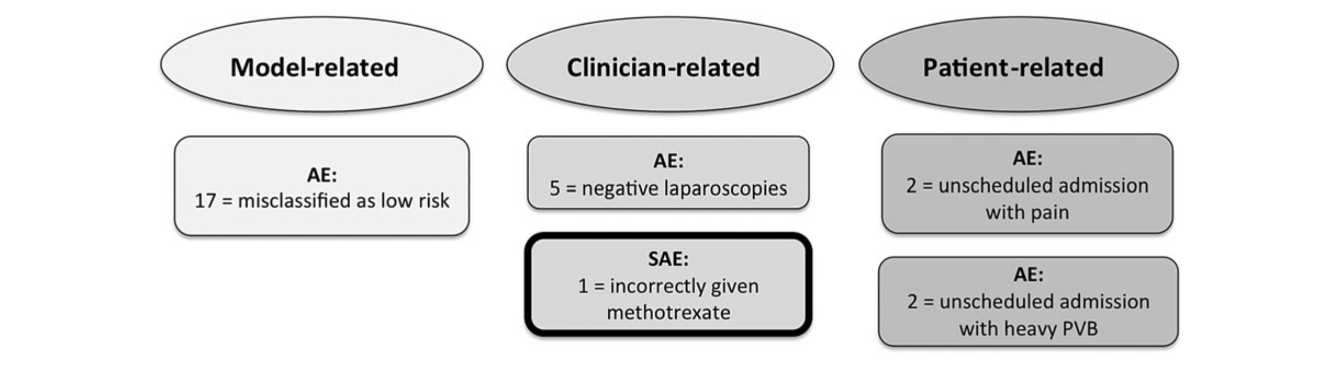
Model-related
Clinician-related
Patient-related
AE:
17 = misclassified as low risk
AE:
5 = negative laparoscopies
SAE:
1 = incorrectly given methotrexate
AE:
2 = unscheduled admission with pain
AE:
2 = unscheduled admission with heavy PVB

Figure 4 Breakdown of adverse and serious adverse events in the study cohort. AE, adverse event; SAE, serious adverse event; PPUL, persistent pregnancy of unknown location; PVB, per vaginal bleeding.

underestimated. This is in line with previous studies (Van Calster *et al.*, 2013). See Supplementary data for detailed results.

Discussion

We have shown that the M4 risk prediction model can be used to triage women with a PUL as being at low or high risk of complications with few AEs. We have further demonstrated that the performance characteristics of the model are retained on a further relatively large patient population when used prospectively in everyday clinical practice by a range of clinicians and allows the follow up of 70% PUL to be minimized with a high NPV for the presence of EP.

A strength of the study is that it was performed in multiple centres with varied population cohorts and staffing levels. Two of the hospitals were university teaching hospitals with clinical research fellows and more than one consultant. The third unit was a district general hospital with a nurse specialist, one consultant and rotating middle grade doctors in attendance. Accordingly, we think these results should be generalizable. Limitations of the study are that 69 (7%) of our PUL were lost to follow-up. Forty-eight hour hCG levels were also missing for 109 (11%) patients, although 68 of these had initial hCG levels <50 with negative urinary hCG tests 2 weeks later, so we know these cases had a final outcome of FPUL.

We also assessed the safety of using the M4 model clinically in two ways: first we described cases where the management recommended by the M4 model was not followed (major and minor protocol deviations) and second we considered model performance where incorrect predictions were made (AE and SAEs).

Where there were minor protocol deviations, 30% (38/125) were the result of doctors overruling the follow-up suggested by the M4 model. This is reassuring as prediction models should be used as a decision support tool rather than as a replacement for clinical judgment. It would be concerning if there were no clinician-related deviations, although when clinicians did not follow the model, there was a tendency to intervene unnecessarily. Nineteen deviations were secondary to unscheduled attendances by participants for increased vaginal bleeding, abdominal pain or anxiety. This is also reassuring as all women classified with a PUL should be warned of the symptoms of an EP and advised to come for review if these occur. In eight of these nineteen cases the final outcome was an EP/PPUL and six of these cases had been incorrectly classified as low risk by the M4 model. No management algorithm for

PUL will ever be perfect as the hCG trend in 10% of EP will mimic that of a viable IUP and 20% will mimic an FPUL (Kirk *et al.*, 2008, 2009, 2014). What these findings show is that as long as women are appropriately counseled to re-present to hospital should they have increased pain/vaginal bleeding, an appropriately timed diagnosis and intervention can still occur without significant harm.

The majority (68/125, 54%) of deviations from the recommended management protocol were secondary to timing issues with blood tests or scans. This can be a practical problem in service provision as in the UK many units that care for women with early pregnancy problems are only open during weekday working hours, or because women simply cannot attend at the required time. We also assessed the total number of blood tests and ultrasound scans required before the final outcome was known (Table IV). As one would expect, a FPUL required the least number of scans (mean $n = 1.1$) whilst EP required the most ($n = 2.8$). For the number of blood tests required, both FPUL and IUP required an average of 2.2, whereas EP required 3.5. These results suggest the M4 model may significantly rationalize the follow-up of low risk PUL.

Table IV **Total number of blood tests and scans required prior to the final diagnosis.**

Deviation	Final diagnosis	Number of scans (SD)	Number of blood tests (SD)
All patients (*n* = 835)	FPUL	1.1 (0.47)	2.2 (0.61)
	IUP	2.4 (0.75)	2.2 (0.53)
	Ectopic	2.8 (1.31)	3.5 (1.97)
None (*n* = 710)	FPUL	1.1 (0.44)	2.1 (0.55)
	IUP	2.4 (0.76)	2.2 (0.53)
	Ectopic	2.7 (1.20)	3.6 (2.04)
Any (*n* = 125)	FPUL	1.3 (0.61)	2.5 (0.85)
	IUP	2.3 (0.65)	2.2 (0.55)
	Ectopic	3.0 (1.62)	3.4 (1.80)
1A (clinician reason) (*n* = 19)	FPUL	1.6 (0.55)	2.8 (0.96)
	IUP	2.5 (1.22)	2.0 (0.00)
	Ectopic	3.0 (0.76)	3.8 (0.71)
1B (patient reason) (*n* = 38)	FPUL	1.3 (0.48)	2.7 (0.67)
	IUP	2.1 (0.50)	2.2 (0.68)
	Ectopic	2.9 (1.51)	2.9 (2.12)
1C (timing) (*n* = 68)	FPUL	1.2 (0.63)	2.4 (0.88)
	IUP	2.4 (0.49)	2.2 (0.54)
	Ectopic	3.3 (3.20)	4.0 (2.45)

Patients are split into PUL that had no protocol deviation, any protocol deviation, deviation 1A (unscheduled visit for a doctor reason), deviation 1B (unscheduled visit for a patient reason), deviation 1C (difference in the timing of a blood test/ ultrasound scan).

Our analysis of AEs shows that although 17 EP/PPUL were misclassified as low risk by the M4 model, none came to significant harm. There were no ruptured EPs, none required a blood transfusion and no participant required a prolonged hospital stay. Overall two women in our study population (1:511), having undergone triage were readmitted and required surgery because of pain. When counseling women about risk this would be classified as an uncommon event (between 1:100 and 1:1000), and may be useful information when advising women on the risks associated with the expectant management of PUL using the M4 model. In both of these cases, no EP was found at laparoscopy.

A number of women in the FPUL group will have had an EP that was not visualized but resolved without intervention—although we do not know which ones these are. Classifying high-risk PUL as those needing intervention rather than solely classing all probable EP as 'high-risk' is an alternative strategy. However, problems lie with consistently applying rules by which clinicians might judge the 'need' to intervene and knowing which EP are 'low-risk', given that rupture can occur even with declining hCG levels. There are data to suggest that the expectant management is appropriate with low serum hCG concentrations (Jurkovic, 2010) and up to 60% of women with an EP/PUL can be managed this way without any significant complications (van Mello *et al.*, 2013). In our study, whilst the model gave management guidance for appropriate follow-up, it was left to individual clinicians to decide about intervention in the event of a probable EP. Similarly, it remains difficult to determine which women with a persistent PUL require intervention.

There was one SAE in the study. This participant was given methotrexate out of hours based on an ultrasound scan suggesting an adnexal mass, and two hCG levels erroneously taken 24, rather than 48, hours apart. This led to the M4 model classifying the PUL as at 'high-risk'. This case had a final outcome of an IUP. After extensive counseling, this pregnancy

Table V **The performance of (i) M4 triage, (ii) triage based on hCG ratio cut-offs (low risk if hCG ratio <0.87 or >1.66) and (iii) triage based on a single progesterone cut-off (low risk if progesterone ≤10 nmol/l).**

	PUL classified as low risk	NPV	Sensitivity	FPR
Using all data (*n* = 835)				
M4 decision	70% (67 to 73)	97% (95 to 98)	82% (73 to 88)	23% (20 to 27)
hCG ratio cut-offs	81% (78 to 83)	96% (94 to 97)	68% (58 to 77)	13% (11 to 15)
Difference	−11% (−13 to −9)	nc	14% (7 to 21)	10% (8 to 13)
Using QCCH and HH patients with progesterone levels (*n* = 560)				
M4 decision	68% (64 to 72)	97% (94 to 98)	82% (72 to 89)	25% (21 to 29)
Progesterone cut-off	47% (43 to 51)	91% (87 to 94)	68% (57 to 78)	51% (46 to 55)
Difference	21% (16 to 26)	nc	14% (−1 to 28)	−26% (−31 to −21)

95% confidence intervals are shown in parentheses.
NPV, negative predictive value; FPR, false positive rate; QCCH, Queen Charlottes' & Chelsea Hospital; HH, Hillingdon Hospital; nc, not computed.

was terminated due to the potential impact of methotrexate on the developing embryo. This critical error highlights the importance of ruling out an IUP before treating with methotrexate.

The management of women with a PUL has shifted towards effective triage rather than identifying pregnancy location. Our study has demonstrated that clinicians will comply with using a prediction model to guide the management of women with a PUL and so rationalize follow-up. This study also assessed the clinical safety of the model with no SAEs associated with correct use of the M4 model in over one thousand PUL. The M4 model is a user-friendly predictive tool that can be used to guide doctors, nurses and midwives when reaching decisions about the management of PUL. It can be accessed at no charge via this link: earlypregnancycare.com/m4triage/.

Supplementary data

Supplementary data are available at http://humrep.oxfordjournals.org/.

Authors' roles

B.V.C., D.T. and T.B. participated in the conception and design of the study. S.B., S.G., F.A., N.M.-J., M.A.-M., H.M., C.S., C.B. and A.K. acquired patient data. B.V.C. performed the statistical analysis. B.V.C., S.B., S.G., D.T. and T.B. interpreted the results. S.B., S.G., B.V.C., D.T. and T.B. wrote the initial version of the manuscript. All authors critically revised the manuscript and approved the final version.

Funding

B.V.C. was partly supported by a postdoctoral fellowship from the Research Foundation Flanders (FWO). T.B. is supported by the NIHR (National Institute of Health Research) Biomedical Research Center based at Imperial College Healthcare NHS Trust and Imperial College London. S.B. is supported by NIHR CLAHRC NWL (Collaboration for Leadership in Applied Health Research & Care, NorthWest London). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. D.T. is Fundamental Clinical Researcher of FWO (Research Foundation – Flanders).

Conflict of interest

None declared.

References

Banerjee S, Aslam N, Zosmer N, Woelfer B, Jurkovic D. The expectant management of women with early pregnancy of unknown location. *Ultrasound Obstet Gynecol* 1999;**14**:231–236.

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

Barnhart K, van Mello NM, Bourne T, Kirk E, Van Calster B, Bottomley C, Chung K, Condous G, Goldstein S, Hajenius PJ *et al.* Pregnancy of unknown location: a consensus statement of nomenclature, definitions, and outcome. *Fertil Steril* 2011;**95**:857–866.

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.

Bottomley C, Van Belle V, Mukri F, Kirk E, Van Huffel S, Timmerman D, Bourne T. The optimal timing of an ultrasound scan to assess the location and viability of an early pregnancy. *Hum Reprod* 2009;**24**:1811–1817.

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 S, 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, Okaro E, Khalid A, Lu C, Van Huffel S, Timmerman D, Bourne T. The accuracy of transvaginal sonography for the diagnosis of ectopic pregnancy prior to surgery. *Hum Reprod* 2005;**20**:1404–1409.

Condous G, Kirk E, Van Calster B, Van Huffel S, Timmerman D, Bourne T. Failing pregnancies of unknown location: a prospective evaluation of the human chorionic gonadotrophin ratio. *BJOG* 2006;**113**:521–527.

Condous G, Van Calster B, Kirk E, Timmerman D, Van Huffel S, Bourne T. Prospective cross-validation of three methods of predicting failing pregnancies of unknown location. *Hum Reprod* 2007a;**22**:1156–1160.

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.

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* 2007c;**29**:680–687.

Cordina M, Schramm-Gajraj K, Ross JA, Lautman K, Jurkovic D. Introduction of a single visit protocol in the management of selected patients with pregnancy of unknown location: a prospective study. *BJOG* 2011;**118**:693–697.

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.

Guha S, Ayim F, Ludlow J, Sayasneh A, Condous G, Kirk E, Stalder C, Timmerman D, Bourne T, Van Calster B. Triaging pregnancies of unknown location: the performance of protocols based on single serum progesterone or repeated serum hCG levels. *Hum Reprod* 2014;**29**:938–945.

Hahlin M, Thorburn J, Bryman I. The expectant management of early pregnancies of uncertain site. *Hum Reprod* 1995;**10**:1223–1227.

Jurkovic D. hCG as a patient. *Ultrasound Obstet Gynecol* 2010;**36**:395–399.

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* 2007;**22**:2824–2828.

Kirk E, Daemen A, Papageorghiou AT, Bottomley C, Condous G, De Moor B, Timmerman D, Bourne T. Why are some ectopic pregnancies characterized as pregnancies of unknown location at the initial transvaginal ultrasound examination? *Acta Obstet Gynecol Scand* 2008;**87**:1150–1154.

Kirk E, Condous G, Bourne T. Pregnancies of unknown location. *Best Pract Res Clin Obstet Gynaecol* 2009;**23**:493–499.

Kirk E, Bottomley C, Bourne T. Diagnosing ectopic pregnancy and current concepts in the management of pregnancy of unknown location. *Hum Reprod Update* 2014;**20**:250–261.

Levine D. Ectopic pregnancy. *Radiology* 2007;**245**:385–397.

Medicines and Healthcare products Regulatory Agency (MHRA). *The Good Clinical Practice Guide*. TSO (The Stationery Office), 2014.

Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998a;**17**:857–872.

Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med* 1998b;**17**: 2635–2650.

Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Stat Med* 2012;**31**:2610–2626.

Van Calster B, Abdallah Y, Guha S, Kirk E, Van Hoorde K, Condous G, Preisler J, Hoo W, Stalder C, Bottomley C *et al*. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Hum Reprod* 2013;**28**:609–616.

Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. *Stat Med* 2014;**10**:2585–2596.

van Mello NM, Mol F, Opmeer BC, Ankum WM, Barnhart K, Coomarasamy A, Mol BW, van der Veen F, Hajenius PJ. Diagnostic value of serum hCG on the outcome of pregnancy of unknown location: a systematic review and meta-analysis. *Hum Reprod Update* 2012;**18**:603–617.

van Mello NM, Mol F, Verhoeve HR, van Wely M, Adriaanse AH, Boss EA, Dijkman AB, Bayram N, Emanuel MH, Friederich J *et al*. Methotrexate or expectant management in women with an ectopic pregnancy or pregnancy of unknown location and low serum hCG concentrations? A randomized comparison. *Hum Reprod* 2013;**28**:60–67.