External validation of models and simple scoring systems to predict miscarriage in intrauterine pregnancies of uncertain viability

S. Guha1,2*, V. Van Belle3,4, C. Bottomley5, J. Preisler2, V. Vathanan5, A. Sayasneh2, C. Stalder2, D. Timmerman6, and T. Bourne2,6

1Hillingdon Hospital NHS Trust, Pield Heath Road, Middlesex UB8 3NN, UK 2Queen Charlotte’s & Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, UK 3ESAT-SCD/iMinds-KU Leuven Future Health Department, KU Leuven, Leuven, Belgium 4Department of Mathematics and Statistics, Liverpool John Moores University, Liverpool, UK 5Department of Obstetrics and Gynecology, Chelsea and Westminster NHS Trust, Fulham Road, London, UK 6Department of Obstetrics and Gynecology, University Hospitals Leuven, Campus Gasthuisberg, KU Leuven, Leuven, Belgium

*Correspondence address. 3 Rosebank Cottages, London Road, Loudwater HP10 9TA, UK. Tel: +7963691608; E-mail: sharmistha.guha@gmail.com

Submitted on March 16, 2013; resubmitted on July 13, 2013; accepted on July 22, 2013

STUDY QUESTION: Does a logistic regression model and scoring system to predict viability of an intrauterine pregnancy of uncertain viability (PUV) perform as well in an independent patient group as the original patient group?

SUMMARY ANSWER: The model and scoring system showed good performance on external validation confirming their value for the prediction of miscarriage/viability in PUV patients up to 11–14 weeks of gestation.

WHAT IS KNOWN ALREADY: Several individual ultrasound and demographic factors have been described as predictors for miscarriage. A logistic regression model and simple scoring system using basic clinical and ultrasound features, such as maternal age, bleeding score, mean gestational sac diameter (MSD) and presence or absence of yolk sac, have been developed to allow patient-specific prediction of viability of PUV beyond the first trimester.

STUDY DESIGN, SIZE, DURATION: Prospective observational external validation cohort study in two inner city early pregnancy assessment units over a period of 18 months.

PARTICIPANTS/MATERIALS, SETTING, METHODS: All consecutive women with a PUV were recruited. Ultrasound (mean sac diameter and presence of yolk sac) and demographic variables (maternal age, bleeding score and gestational age) were noted. The outcome measure was first trimester (11–14 week) viability. Women with unknown first trimester outcome were excluded. Receiver operating characteristic (ROC) curves and calibration plots were constructed. Test performance was compared with the original development data set. A new model and scoring system, which did not include gestational age, was built and evaluated.

MAIN RESULTS AND THE ROLE OF CHANCE: Of the 575 women who were recruited, first trimester outcome was known for 89.2% (n = 513). The model could only be validated in 400 patients, due to missing values in model variables and outcome. The model predicted viability with an area under the ROC curve (AUC) of 0.845 [95% confidence interval (CI), 0.806–0.884] compared with 0.774 (95% CI, 0.701–0.848) in the original study. The AUC for the scoring system was 0.832 (95% CI, 0.792–0.872) compared with 0.771 (95% CI, 0.698–0.844) from the original study data set. The new model and the scoring system, excluding gestational age, could be evaluated on 503 patients and resulted in an AUC of 0.801 (95% CI, 0.765–0.841) for the model and 0.773 (95% CI, 0.733–0.812) for the scoring system.

LIMITATIONS, REASONS FOR CAUTION: Approximately 22% of patients could not be validated due to missing variables and for 11% of patients the first trimester outcome was unknown.

WIDER IMPLICATIONS OF THE FINDINGS: Both the model and the scoring system showed excellent performance on external validation confirming their generalizability and utility in prediction of viability beyond the first trimester in clinical practice. An advantage of the mathematical models original Mo and new Mn and scoring systems original SS0 and new SSn is that they can provide women with an individualized...
probability of the viability of their pregnancy using only demographic information, symptoms and TVS findings. Furthermore, the risk of miscarriage can be given immediately following examination.

**STUDY FUNDING/COMPETING INTEREST(S):** T.B. is supported by the Imperial Healthcare NHS Trust NIHR Biomedical Research Centre. This research is supported by Research Council KUL GOA MaNet, iMinds 2012, Belgian Federal Science Policy Office IUAPP719. VVB is a postdoctoral fellow of the Research Foundation — Flanders (FWO). There are no conflicts of interest.

**Key words:** gestational sac / miscarriage / prediction models / pregnancy of uncertain viability / ultrasound

**Introduction**

In a systematic review of the accuracy of first trimester ultrasound for the diagnosis of early embryonic demise, Jeve et al. (2011) found that there was a paucity of high-quality prospective data on which to base guidance on diagnosing miscarriage. Three further recent publications suggested that criteria used to define miscarriage based on transvaginal scan (TVS) measurements of gestation sac and embryo size were unreliable (Abdallah et al., 2011a,b; Pexsters et al., 2011). This new information led to the Royal College of Obstetricians and Gynaecologists (RCOG) in the UK amending their guidance such that on a single TVS a miscarriage can only be diagnosed when an empty gestation sac is ≥25 mm in mean diameter or when the crown-rump length (CRL) measurement for an embryo with no visible heartbeat is ≥7.0 mm. The guidance emphasized the need for repeating ultrasound scans at an interval in order to definitively comment on viability (RCOG Green-top Guideline, 2006, Addendum to RCOG Green-top Guideline 25, 2011). At a consensus meeting of the Society of Radiologists in Ultrasound in 2012 it was concluded that similar guidance should be adopted in the USA. Whilst this more conservative approach to the diagnosis of miscarriage is welcomed, it is likely to be associated with more pregnancies being classified as being of uncertain viability. The introduction of sensitive home pregnancy tests means that women may have a positive pregnancy test even before they have missed a period. The result of these developments means that women attending for TVS in early pregnancy are more likely to have an uncertain outcome and be asked to undergo a repeat TVS at intervals of 7–14 days.

We know that early pregnancy complications may lead to significant psychological morbidity (Lok and Neugebauer, 2007). The uncertainty of waiting for repeat examinations may compound what is already an anxious situation for couples and, furthermore, there is often a reluctance to accept that definitive answers cannot always be given at a single visit. It is helpful to counsel women about the probable outcome of the pregnancy so that they are better prepared for the likely findings when they return to the clinic.

Anxiety also arises when a pregnancy is found to be smaller than expected for the menstrual dates given by the patient. This may be an innocent finding and relate to variations in both the timing of ovulation and implantation (Mahendru et al., 2012). However, it has been appreciated that a difference in embryonic size from that expected may reflect a higher risk of miscarriage or aneuploidy (Mukri et al., 2008). For other women the discrepancy between expected and observed embryo measurements may be because they have irregular periods or do not accurately remember the date of their last menstrual period (LMP). When giving appropriate advice about the possible outcome of pregnancy to women in these circumstances, it is important to avoid unrealistic expectations.

To address some of these issues we have previously described a mathematical model (M<sub>n</sub>) and simple scoring system (SS<sub>n</sub>) to predict the outcome of a pregnancy when it is classified as an intrauterine pregnancy of uncertain viability (PUV) (Bottomley et al., 2011). Both the model (M<sub>n</sub>) and the scoring system (SS<sub>n</sub>) provide an individualized probability of pregnancy viability depending on maternal age, the amount of vaginal bleeding, gestational age, the mean gestational sac diameter and visibility of a yolk sac. The M<sub>n</sub> and SS<sub>n</sub> were developed on data from one hospital unit, and showed very good prediction of viability on internal validation. However, before introducing a test into routine clinical practice it is necessary to demonstrate a good performance in different settings and populations.

In the present study we carried out an external validation to test the performance of both the M<sub>n</sub> and SS<sub>n</sub>. The two units are inner London hospitals with large clinical throughputs and ethnically diverse populations. We also approached the problem of how to include women who are uncertain about the date of their LMP in the prediction model. We therefore used the training set from the original paper (Bottomley et al., 2011) to develop a new model (M<sub>n</sub>) and scoring system (SS<sub>n</sub>) that did not take gestational age into account. This model and the scoring system were then tested on both the original and external validation data.

**Materials and Methods**

This was an 18-month multicentre prospective cohort study (January 2011-July 2012) in the early pregnancy assessment units of Chelsea and Westminster Hospital (C&W) and Queen Charlottes and Chelsea Hospital, Imperial College, London (QCCH). Both units take referrals from primary care physicians, emergency units and other departments in the hospital. There is no minimum gestational age criteria for attendance.

Inclusion criteria for the study were: all women classified as having a PUV. A PUV was defined using the following criteria on the basis of a TVS:

- The visualization of a single intrauterine gestational sac with an MSD of ≤20 mm at C&W (prior to September 2011) and ≤25 mm at QCCH, with or without a visible yolk sac but without a visible embryo, or a single intrauterine gestational sac with a visible embryo of ≤6 mm at C&W (prior to September 2011) and ≤7 mm at QCCH but no embryonic heart beat. Exclusion criteria were women with multiple pregnancies and those who underwent subsequent termination of pregnancy.

Prior to the consultation, the women completed a written questionnaire which included the date of their LMP, previous obstetric history and pregnancy symptoms. This information was confirmed by the examining sonographer. The amount of vaginal bleeding was estimated using a pictorial bleeding assessment chart (PBAC) (Higham et al., 1990). The bleeding score ranged from 0 (no bleeding) to 4 (bleeding with clots).

All women underwent a TVS performed by an appropriately trained examiner (a nurse specialist, gynaecologist or sonographer) using either a Voluson® E8 Expert (GE Healthcare, Wisconsin, USA) or Medison Accuvix® XG (Samsung Medison, Seoul, South Korea) ultrasound machine using a tight curvilinear transvaginal probe operating at a frequency of...
3.5–9.3 MHz. Structured assessment consisted of measurements of the gestational sac in three orthogonal planes and measurement of the length of the embryo if present. All images were stored in the Picture Archiving and Communication System or as hard copies in the patient’s casenotes. The measurements were reported using commercially available ultrasound databases (Radcentre, iSOFT; IBA Healthgroup Company) and Astraia (Astraia software GMBH, Munich, Germany).

All women included were scheduled to have a repeat TVS examination after 7–14 days. The outcome of interest was viability at the routine end of first trimester scan (11–14 weeks of gestation).

The study did not need ethical approval and was classed as an audit. Approval was obtained from the local Institute Review Board.

**Statistical analysis**

In our validation study, only the five variables included in the previously developed model were recorded. In the original model, univariable and multivariable logistic regression analyses had been performed to establish the relationship between first trimester viability and a number of variables including maternal age, ethnicity, obstetric history, gestational age by LMP, abdominal pain, pain score, presence of vaginal bleeding, PBAC, mean sac diameter, presence of yolk sac, mean yolk sac diameter and subchorionic haematoma. Multivariable models were developed with the significant variables, using a stepwise approach (Sullivan et al., 2004; Bottomley et al., 2011). These were maternal age, gestational age (calculated from LMP), PBAC, MSD and presence or absence of a yolk sac. For both the mathematical model ($M_a$) and simple scoring system ($SS_n$), a receiver operating characteristic (ROC) curve was constructed to describe the relationship between the sensitivity and the false-positive rate (1 – specificity) to predict ongoing viability. Calibration plots, plotting the observed versus the predicted probability of viability for both the mathematical model and the scoring system, were constructed.

As the $M_a$ and $SS_n$ could not be used in women with unknown LMP, a new mathematical model ($M_n$) and scoring system ($SS_n$) were developed on the...
training set of the original cohort. These were dependent on four variables: maternal age, PBAC, mean sac diameter and presence or absence of a yolk sac. The new $M_n$ and $SS_n$ were applied to the original test set and further tested on the new external validation data set.

Results

During the study period 575 consecutive women with a PUV were recruited. The outcome was known in 513 (89.2%) pregnancies, and 260/513 (50.7%) pregnancies were viable and 253/513 (49.3%) non-viable at the end of the first trimester. Analysis was performed on the 400 patients for whom all covariates and the outcome were known. Most exclusions ($n = 103$) were due to unknown LMP. The PBAC and presence of the yolk sac was unknown for nine and two pregnancies, respectively, whereas one patient had both variables missing. Of the 400 pregnancies, 200 (50%) were viable and the rest were non-viable at the end of first trimester. The data collection method is shown in Fig. 1. Patients who miscarried had heavy bleeding in the interim with no pregnancy visualized on the follow-up scan at 7–14 days, a non-viable pregnancy diagnosed at a 7–14 day follow-up scan or had a viable pregnancy at the follow-up scan but subsequently miscarried before the end of the first trimester. The data collection method is shown in Fig. 1. Patients who miscarried had heavy bleeding in the interim with no pregnancy visualized on the follow-up scan at 7–14 days, a non-viable pregnancy diagnosed at a 7–14 day follow-up scan or had a viable pregnancy at the follow-up scan but subsequently miscarried before the end of the first trimester.

The model ($M_o$) was defined as

$$z = 9 - 0.27 \times \text{maternal age - 0.4} \times \text{PBAC score} + 1.3 \times (\text{gestational age < 42 days})$$

$$- 0.09 \times (|\text{MSD} - 7|) + 1.3 \times (\text{yolk sac present}),$$

where maternal age is taken to be 35 years when the actual age is >35 years. The estimated chance of a viable pregnancy is then obtained as $\exp(z)/(1 + \exp(z))$. On the new external validation set, $M_o$ obtained an area under the ROC curve (AUC) of 0.845 (95% confidence interval (CI) 0.806–0.884). The $SS_o$ obtained an AUC of 0.832 (95% CI 0.792–0.872) in the same set. The ROC curves of both the model and the scoring system are illustrated in Fig. 2.

As described above, the original model is not able to obtain a risk estimate for the significant number of women who are unable to report the LMP date with accuracy. The new model ($M_n$) and scoring system ($SS_n$) that did not include information on LMP, were thus built using the training set from the original study data set (Bottomley et al., 2011). The resulting model is defined as

$$z = 10.87 - 0.29 \times \text{maternal age - 1.03} \times \text{PBAC score}$$

$$- 0.13 \times (|\text{MSD} - 7|) + 0.77 \times (\text{yolk sac present}) + 0.91 \times \text{PBAC score} \times (\text{yolk sac present}),$$

where maternal age was again taken to be 35 years when actual age is >35 years. This model was validated on the 503 patients with known outcome in the external validation study. The model obtained an AUC of 0.801 (95% CI 0.765–0.841).

Table I summarizes the new score system $SS_n$. The score system obtained an AUC of 0.773 (95% CI 0.733–0.812) on the new external validation data. The ROC curves for the new model and scoring system are given in Fig. 3. Calibration plots (not shown) illustrate that the model and scoring systems are well calibrated, indicating that the chance predicted by the model corresponds to the chance observed in the data set.

Discussion

We have shown that a mathematical model developed to predict the outcome of pregnancies of initially unknown viability maintains its performance when subjected to external validation. We have also demonstrated that a new model that does not require knowledge of LMP as a variable has reasonable test performance. This is important as our study showed that 20% of women with a PUV cannot accurately recall the date of their LMP.

In comparison with the original study, the external validation data set showed a similar miscarriage rate (50 versus 50.7%, respectively). Approximately 39.8% women presented with symptoms of bleeding compared with 35.5% in the original population. Both the development and external validation studies took place in ethnically diverse busy inner city hospitals.

Several tools have been described in the literature for predicting the outcome of pregnancies of uncertain viability; however, these usually require information derived from both TVS and biochemical studies of blood. Bignardi et al. (2008) found that an hCG ratio $> 2.0$ was predictive of a viable pregnancy at 11–14 weeks. However, this study required two blood tests and was based on a population of women initially classified with pregnancies of unknown location. Another logistic regression model was originally published by Elson et al. (2003) and was dependent not only on maternal demographics and ultrasound features but also on serum progesterone measurement. This model showed reasonable performance on temporal validation with an AUC of 0.85 compared with
0.97 in the original study (Lautmann et al., 2011). However, recruitment for the validation study took 6 years because over 90% of eligible women refused the test, possibly because of the necessity for blood tests. In routine clinical practice, hCG analysis is not performed for women who are diagnosed with a PUV at the initial TVS. Therefore, our study was deliberately based on simple indicators that would be normally ascertained as part of routine clinical practice.

An advantage of the mathematical models $M_o$ and $M_n$ and scoring systems $SS_o$ and $SS_n$ validated in this paper, is that they can provide women with an individualized probability of the viability of their pregnancy at the end of the first trimester, only using demographic information, symptoms and TVS findings. Furthermore, the risk of miscarriage can be given immediately following examination. For example, a 35-year old woman with a bleeding score of 2–48 days of gestation with a mean gestational sac diameter of 14 mm and absent yolk sac has an 18% probability of viability at the end of first trimester using the original scoring system $SS_o$ (or 13.2% using the new $SS_n$ that does not take gestational age into account).

The new $M_n$ and $SS_n$, which do not require LMP as a variable, did not perform as well as the original $M_o$ and $SS_o$. However, they showed reasonable test performance which was retained on external validation, and so give us a tool which we can use in the significant minority of women with uncertain dates.

The scoring system is easy to use and does not require a computer for use (see Fig. 4). The model and scoring system can be integrated into any of the commercially available ultrasound reporting programmes or into a smartphone or tablet application, and so their introduction into clinical practice should be straightforward. The models could also be developed into a mobile phone application or software (Van Belle et al., 2012).

At present, a multicentre prospective study is ongoing to better define the cut-off levels for measurements of MSD and CRL that can be used to diagnose miscarriage (Diagnosis of Miscarriage Study). It is likely that this would support recommendations that more women are asked to return for a repeat scan to definitively check viability. A reliable estimated probability of the likely outcome of their pregnancy would be helpful when counselling these women. In the event of a high likelihood of a viable pregnancy, women could be reassured that they do not need close follow-up (Westin et al., 2007). When the predicted outcome is a non-viable pregnancy, women could be warned regarding possible heavy bleeding and offered a repeat TVS assessment after an interval of 7–14 days.

A very important issue in early pregnancy care is managing the realistic expectations and anxieties of women who present with bleeding. Our...
experience is that when given a choice, women prefer to be given an immediate individualized probability of the likely outcome of their pregnancies rather than being given the standard advice that there is a background 50% chance of miscarriage for all women with a PUV. A study using a predictive model and scoring system to predict the chance of viability in all women seen in an early pregnancy assessment unit (including those with an embryonic heart beat at initial TVS) is being validated at other units to assess generalizability and patient acceptability (Bottomley et al., 2013).

We have shown that a prediction model or scoring system to predict viability works in different clinical settings. These prediction tools can be introduced into clinical practice to both counsel women about the likely outcome of their pregnancy, and offer guidance on appropriate follow-up arrangements based on the prognosis for that particular pregnancy.

**Authors’ roles**

C.B. conceptualized the idea of the article. S.G. and C.B. designed the study. S.G., V.V., J.P. collected the data. V.V.B. performed the statistical analysis. S.G. drafted the manuscript. V.V.B., C.B., A.S., C.S. and D.T. contributed to writing of the article. T.B. had the original idea for the article, participated in data analysis, decision-making and supervised writing of the article.

**Funding**

T.B. is supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This research is supported by Research Council KUL GOA MaNet, iMinds 2012, Belgian Federal Science Policy Office IUAP P719. V.V.B. is a postdoctoral fellow of the Research Foundation—Flanders (FWO).

**Conflict of interest**

None.
References


