Diagnostic research in perspective: examples of retrieval, synthesis and analysis

Bachmann, L.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Multivariable analysis of tests for the diagnosis of intrauterine growth restriction

Lucas M. Bachmann, Khalid S. Khan, Joseph Ogah, Philip Owen

Ultrasound Obstet Gynecol. 2003 Apr;21(4):370
Abstract

Design: Prospective cohort study

Subjects: 274 low-risk pregnancies undergoing serial ultrasound examination at pre-determined intervals:

Objectives: 1. To describe how data from antenatal fetal ultrasound biometry, amniotic fluid index and umbilical artery Doppler can be appropriately combined using multivariable models. 2. To investigate how the addition of these ultrasound parameters influences the ability to predict intrauterine growth restriction (IUGR)

Methods: Standard deviation (Z) scores of the last values for Fetal Abdominal Area (FAA), growth velocity of the FAA (FAA vel), amniotic fluid index (AFI) and umbilical artery Doppler pulsatility index prior to delivery were calculated for 260 fetuses. Customised estimated fetal weight (cEFW) centiles were also calculated using the last EFW before delivery after adjustment for fetal gender, gestational age, birth order, maternal weight, height and ethnic origin. Following delivery the neonatal ponderal index was calculated and centile position obtained. Neonatal ponderal index below the 25th centile served as the main outcome measure for diagnosis of IUGR. Logistic regression analysis was used to delineate the predictive value of the three fetal growth tests FAA, FAA velocity and cEFW and the additional value of AFI and pulsatility index of the umbilical artery.

Results: The areas under the ROC curves (95% Confidence Interval) for FAA, FAA vel and cEFW alone were 0.819 (0.748-0.891), 0.784 (0.699-0.869) and 0.74 (0.643-0.837) respectively in the prediction of a neonatal ponderal index <25th percentile. The addition of both the AFI and pulsatility index to FAA, FAA vel and cEFW generated small increases in the areas to 0.831 (0.758-0.904), 0.817 (0.735-0.899) and 0.766 (0.672-0.859) respectively. These improvements in diagnostic prediction were not statistically significant.

Conclusions: The addition of AFI and umbilical artery pulsatility index to the fetal biometry parameters did not significantly increase the ROC areas in the study population. The approach applied in this study is useful in the context of hypothesis generation. Further studies using larger data sets and other predictors should be carried out using the analytical techniques outlined in this paper to determine the contribution of various antenatal tests in the prediction of IUGR.
Introduction
Fetal assessment with ultrasound biometry, amniotic fluid volume assessment and umbilical artery Doppler waveform analysis is frequently practised in modern antenatal care. When the clinician is presented with a number of ultrasound parameters, decision-making is inevitably based upon a clinically intuitive combination of these tests. In the published literature, tests are evaluated for their performance independently of each other but such an approach is associated with bias.

Fetal biometry expressed either as a single measurement or as change in serial measurements performs with varying degrees of success in the prediction of intra-uterine growth restriction (IUGR) in both low and high-risk pregnancies. To our knowledge there is no published description of an appropriate statistical methodology for combining information from more than one ultrasound parameter in the prediction of infants born with evidence of IUGR. This study describes how several ultrasound fetal variables can be appropriately combined using multivariable models and how the addition of different ultrasound parameters influences test performance. In particular, we wanted to examine if the addition of amniotic fluid index (AFI) and Doppler examination of the umbilical artery to standard and customised fetal biometry improved diagnostic prediction.

Methods
Three hundred and thirteen women attending the antenatal clinic at Ninewells Hospital, Dundee, Scotland were enrolled into a study of fetal growth which has previously been described in detail. Entry criteria were singleton pregnancy, gestational age of less than 85 days confirmed by crown-rump length measurement and the absence of recognised risk factors for accelerated or retarded fetal growth including a history of a previous small for gestational age infant, existing medical disorders or heavy smoking (>20 cigarettes per day). All the subjects were scanned for fetal anomaly at 18 weeks and subsequently underwent ultrasound examinations at 4 week intervals before 30 weeks and 2 week intervals thereafter until delivery.

All ultrasound measurements were made using an Aloka SSD-650 real-time ultrasound scanner using a 3.5 MHz probe by one of the authors (PO). Crown-Rump Length (CRL) was measured in a standard manner employing a frozen on-screen image and electronic calipers. Gestational age was calculated with reference to CRL and not menstrual data. The Fetal Abdominal Area (FAA) was measured at the level of the umbilical vein by tracing the outline of the trunk on-screen. For each fetus the last values for FAA prior to delivery were converted to standard deviation (Z) scores using previously published gestational age specific values for means and standard deviations. Growth velocity of the FAA measurement was determined from the last and third last FAA measurements for each fetus; FAA velocity was calculated over this mean 28 day interval and expressed as a Z score using previously published standards.
Estimated fetal weight (EFW) was calculated from bi-parietal diameter (BPD), FAA and femur length (FL) measurements using a previously validated formula. A customised estimated fetal weight (cEFW) centile was calculated for each fetus using the last EFW before delivery after adjustment for fetal gender, gestational age, maternal weight at booking, birth order and maternal ethnic origin; this was facilitated by on-line software available at www.wmpi.net. FAA Z score, FAA velocity Z score and cEFW served as the measures of fetal size and growth.

Amniotic fluid index was measured once at each visit using a standard technique. The umbilical artery pulsatility index (PI) was calculated from the mean of three waveforms obtained by insonating a free-floating segment of cord during fetal quiescence and in the absence of fetal breathing. The last values for AFI and PI prior to delivery are described as Z-scores.

Neonatal anthropometric measurements were made on the second or third days of life. The baby's length was measured on a standard neonatal anthropometer; the mean of three measurements was recorded, the Ponderal index calculated and centile position obtained (PI = weight/crown-heel length$^3$ (g/cm$^3$) X 100). In addition, neonatal skinfold thickness was measured using Holtain calipers (Crymych, UK) by one observer. Three measurements were made at the subscapular and triceps areas on the child, the mean measurement recorded and a centile position obtained after adjustment for gestational age and sex.

Infants were considered to have IUGR if the ponderal index was below the 25$^{th}$ centile and/or one or both skinfold thickness measurements were below the 10$^{th}$ centile. These cut-off values were chosen because they identify only a small proportion of the low-risk population as being growth restricted and as such are likely to represent true growth restriction. There is no consensus as to which neonatal measurement best represents nutritional status but ponderal index is more widely reported than skinfold thickness. For this reason, the PI was selected as the main outcome measure (results for skinfold thickness are presented in the appendix).

Statistical evaluation of combination of diagnostic tests

Logistic regression analysis was used to delineate the predictive value of the three fetal size and growth tests (FAA, FAA velocity and cEFW) separately and with the additional value of AFI and umbilical artery pulsatility index Z scores. For our analysis we used Ponderal Index below the 25$^{th}$ percentile as the binary dependent (outcome) variable. In the model building process, firstly univariable analyses were carried out for each of the three fetal size and growth tests to predict IUGR. Secondly, each of the three biometry tests was combined with AFI and pulsatility. To assess the stability of our inferences, the analyses were repeated with skinfold thickness <10$^{th}$ percentile as the dependent variable. In total, we ran 24 regression models using the SPSS® statistical software package (Version 10.0 SPSS Inc. 2001. Chicago, Illinois, USA).
The predictive value of each model was initially summarized by plotting the estimates of sensitivity (true positive rates) against 1 - specificity (false positive rates) to develop a receiver operating characteristic (ROC) curve that characterized the performance of the particular diagnostic model \(^{17,18}\). This approach enables comparison of the models by analysing the difference in magnitude of the ROC areas. A difference between areas is then assessed for statistical significance\(^ {19,20,21,22}\). The area under the ROC curve provides information about the entire range of test results. An ROC area of 0.5 describes a non-informative test whereas a ROC area of 1.0 represents a fully informative test that discriminates between disease presence and absence perfectly. The ROC curves generated from logistic regression modelling and their respective areas were compared. This approach allowed an assessment of the incremental value of the addition of AFI and umbilical artery pulsatility index over and above that of fetal size and growth alone.

**Results**

Two hundred and seventy four women continued in the study. A total of 260 (95%) delivered at 37 weeks gestation or more. Twenty-two (8%) and 11 (4%) had adjusted birthweights below the 10\(^{th}\) and third centiles respectively. There were no cases of structural or chromosomal anomaly. Skinfold thickness and ponderal index were available in 238 and 257 cases respectively. Twenty-six (10.9%) infants had one or both skinfold thickness measurements below the tenth centile and 40 (15.6%) had a ponderal index below the 25\(^{th}\) percentile.

Two hundred and fifty eight cases had complete ultrasound and maternal data (FAA, FAA velocity, cEFW, AFI and pulsatility index) with one or both ponderal index and/or skinfold thickness centiles (242 and 226 respectively); it is these 258 cases which forms the population undergoing further analysis. For these 258 cases with complete ultrasound data and at least one measure of growth achievement the mean and range of the Z scores were: FAA -1.8 (0.44 to -3.9), FAA velocity -0.13 (2.87 to -3.1), AFI -0.6 (2.93 to -2.8) and pulsatility index 0.08 (-1.28 to 2.7). The mean and range of cEFW was 39 (0.6 to 99.8).
Fig 1a: Plot of standard deviation (Z) scores for AFI and umbilical artery PI for cases with FAA Z scores >-2.5 (AFI2SDS = Z score for the Amniotic Fluid Index. PI2SDS = Z score for the umbilical artery Doppler PI)

Fig 1b: Plot of standard deviation (Z) scores for AFI and umbilical artery PI for cases with FAA Z scores <-2.5 (AFI2SDS = Z score for the Amniotic Fluid Index. PI2SDS = Z score for the umbilical artery Doppler PI)

The distribution of the values for AFI and umbilical artery Doppler PI for both small (FAA Z score <-2.5) and appropriate sized fetuses (FAA Z score >-2.5) is presented in figures 1a,1b (a FAA Z score of -2.5 was chosen after inspection of the ROC curve). There is a wide range of values for AFI and umbilical artery recorded amongst both small and appropriately sized subjects.

Tests for predicting Ponderal Index <25th centile showed that the areas under the ROC curves for FAA, FAA velocity and cEFW alone were 0.819 (0.748-0.891), 0.784 (0.699-0.869) and 0.74 (0.643-0.837) respectively. The stepwise addition of information from the AFI increased the area only slightly to 0.82 (0.747-0.894), 0.787 (0.702-0.871) and 0.75 (0.652-0.847) respectively. Combining fetal size and growth
Multivariable analysis of tests for the diagnosis of intrauterine growth restriction

measures with the pulsatility index resulted in areas of 0.829 (0.758-0.901), 0.817 (0.734-0.901) and 0.762 (0.669-0.856). The full model consisting of one of the fetal size and growth tests (FAA, FAA velocity, cEFW), AFI and pulsatility index increased the area to 0.831 (0.758-0.904), 0.817 (0.735-0.899), and 0.766 (0.672-0.859). (Table 1) The results for the combination of tests in the prediction of skinfold thickness <10th centile showed a pattern similar to that observed above for the ponderal index (appendix).

Table 1: Accuracy of three ultrasonic fetal biometric tests in the prediction of intrauterine growth restriction and the additional benefit of measuring amniotic fluid index (AFI) and pulsatility index (PI) of the umbilical artery.

<table>
<thead>
<tr>
<th>Neonatal Outcome</th>
<th>Regression models for estimating accuracy of combinations of tests*</th>
<th>Ultrasonic Fetal Biometric Tests</th>
<th>Fetal Abdominal Area</th>
<th>Velocity of Fetal Abdominal Area</th>
<th>Customized Estimated Fetal Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ponderal Index</td>
<td></td>
<td>Fetal Abdominal Area</td>
<td>0.819 (0.748-0.891)</td>
<td>0.784 (0.699-0.869)</td>
<td>0.74 (0.643-0.837)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fetal Abdominal Area + AFI</td>
<td>0.82 (0.747-0.894)</td>
<td>0.787 (0.702-0.871)</td>
<td>0.75 (0.652-0.847)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fetal Abdominal Area + PI</td>
<td>0.829 (0.758-0.901)</td>
<td>0.817 (0.734-0.901)</td>
<td>0.762 (0.669-0.856)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fetal Abdominal Area + AFI + PI</td>
<td>0.831 (0.758-0.904)</td>
<td>0.817 (0.735-0.899)</td>
<td>0.766 (0.672-0.859)</td>
</tr>
</tbody>
</table>

*Accuracy estimation based on logistic regression models using neonatal ponderal index < 25th centile as the outcome variable. Predictor variables included one of the fetal biometry tests with or without AFI and/or umbilical artery PI.

Discussion

This study demonstrates a stepwise, multivariable approach for combining diagnostic tests of fetal growth and fetal wellbeing in the prediction of infants with anthropometric features of intrauterine growth restriction. We selected neonatal anthropometry as our outcome measure rather than birthweight for gestational ages since anthropometric measures correlate more closely with subsequent short and long-term outcomes for the infant23,24.

Previously published univariable analyses have outlined the potential and limitations of several of these parameters in the study population and in other populations2,3,4,25 but until now there has been no published description of the appropriate methodology for combining test performances in the prediction of IUGR. In this study, multivariable analyses, built to reflect the clinical sequence in which a variety of diagnostic tests might be employed showed that the addition of information from amniotic fluid volume evaluation and Doppler examination of the umbilical arteries to any of three measures of fetal biometry only marginally increased the ability to correctly identify IUGR.

We chose to add estimates of amniotic fluid volume and umbilical artery Doppler resistance to fetal biometry since they provide information on different aspects of placental function and pathology26,27. The results of this study should not be interpreted as suggesting that umbilical artery Doppler is not a useful test in the clinical management of the SGA or high-risk pregnancy where the end-point is fetal or neonatal
hypoxia/acidosis. These outcomes are uncommon in low-risk pregnancies and a much larger population would be required to evaluate the addition of umbilical artery Doppler to biometry using the methodology described here.

Currently, diagnostic research tests are often evaluated as if they provide information independently of each other. Estimates of diagnostic accuracy derived in this way ignore information that may have already been obtained from prior testing in the diagnostic process. Such an approach can lead to erroneous inferences and may artificially inflate the value of test combinations.

We further compared the diagnostic probabilities generated by a simple Bayesian approach (assuming complete independence of fetal growth and wellbeing) with those derived from the multivariable model described here (which takes account of the overlap of information between the tests). When employing FAA Z score as the measure of fetal size, figure 2 illustrates that the simple Bayesian approach overestimates the probability of correctly identifying IUGR by almost 20%. This highlights the potential for bias inherent in simple diagnostic analyses without logistic regression modelling. This overestimation of disease probability can have an adverse impact on clinical decision-making, an issue requiring further investigation.

![Fig. 2: Predictions of probabilities for Ponderal Index below the 25th percentile combining information from estimates of fetal size, AFI and umbilical artery Doppler. Comparison of a simple Bayesian approach and logistic regression modeling (see text for details). Optimal cut-off values for a positive test result were chosen using ROC curves (FAA -2.5; AFI -0.8; PI 0.6) \(^{325}\).](image)

Our inferences based on the analysis of ROC areas should be seen in the context of hypothesis generation; this is because a ROC curve is derived using the entire range of possible cut-offs of a positive test result.
In clinical practice tests are usually applied at a specific cut-off for a positive result therefore clinical decisions are only made at specific parts of the ROC curve. In addition, comparisons of diagnostic strategies based on the ROC area can become problematic if the curves differ in their shape and especially if their lines cross 28.

We believe this paper provides guidance for future research involving larger data sets but cannot be used for defining clinical practice at this initial stage. There is currently no consensus on the most appropriate method of sample size calculation for multivariable analyses but a study should aim to have at least 10 cases with the condition per diagnostic variable analysed. For example, if a condition has a prevalence of 10% and five variables or indicators are included a minimum number of 500 cases is required 29. To further guide clinical decision making, evaluations of test performances and the cost-consequences from testing (or not testing) at clinically meaningful treatment thresholds of different diagnostic strategies will be required.

Acknowledgement
P. Owen is grateful to Wellbeing, the charitable arm of the Royal College of Obstetricians and Gynaecologists, London for financial support during the initial collection of fetal biometry.

References


10. Aoki M. Fetal weight calculation; Osaka University Method. *Ultrasound in Obstetrics and Gynaecology ed Ciba, Y Ch 7 pp 95-107 1990*.


15. Lubchenco LO, Hansman C, Boyd E. Intrauterine growth in length and head circumference as estimated from live births at gestational ages from 26 to 42 weeks. *Pediatrics* 1966; 37(3):403-408.


Multivariable analysis of tests for the diagnosis of intrauterine growth restriction


Appendix

Accuracy of three ultrasonic fetal biometric tests for prediction of intrauterine growth restriction and the additional benefit of measuring amniotic fluid index (AFI) and pulsatility index (PI) of the umbilical artery.

<table>
<thead>
<tr>
<th>Neonatal Outcome</th>
<th>Regression models for estimating accuracy of combinations of tests*</th>
<th>Ultrasonic Fetal Biometric Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skinfold thickness</td>
<td></td>
<td>Fetal Abdominal Area</td>
</tr>
<tr>
<td>Fetal biometry test alone</td>
<td>0.802 (0.719-0.885)</td>
<td>0.742 (0.653-0.832)</td>
</tr>
<tr>
<td>Fetal biometry test + AFI</td>
<td>0.83 (0.747-0.912)</td>
<td>0.786 (0.7-0.872)</td>
</tr>
<tr>
<td>Fetal biometry test + PI</td>
<td>0.82 (0.745-0.896)</td>
<td>0.791 (0.702-0.88)</td>
</tr>
<tr>
<td>Fetal biometry test + AFI + PI</td>
<td>0.843 (0.766-0.921)</td>
<td>0.821 (0.742-0.9)</td>
</tr>
</tbody>
</table>

*Accuracy estimation based on logistic regression models using neonatal skin fold thickness < 10th centile as the outcome variable. Predictor variables included one of the fetal biometry tests with or without AFI and umbilical artery PI. Data presented as area under receiver operating characteristic curve (95% Confidence intervals)