DOI:
10.1002/uog.20401

Document Version
Peer reviewed version

29th World Congress on Ultrasound in Obstetrics and Gynecology

12-16 October 2019, Berlin, Germany

Join us in Berlin at the leading event in obstetric and gynecological ultrasound.

Register now

12 August	Early bird registration deadline
12 October	Pre-Congress courses
13 October	Congress opens

For full details visit **isuog.org/worldcongress/2019**

congress@isuog.org | **+44 (0)20 7471 9955**

* Discounts apply to ISUOG members, trainees and sonographers

Organised by the International Society of Ultrasound in Obstetrics and Gynecology.

#ISUOG2019

isuog.org

Development and validation of the predictive models for the QUiPP App v.2: a tool for predicting preterm birth in high-risk asymptomatic women

Dr Helena A. Watson, Mr Paul T. Seed, Dr Jenny Carter, Dr Natasha L. Hezelgrave, Dr Katy Kuhrt, Prof Rachel M. Tribe and Prof. Andrew H. Shennan

Department of Women and Children's Health, School of Life Sciences

King's College London

10th Floor, North Wing

St. Thomas' Hospital

Westminster Bridge Road

LONDON SE1 7EH

Corresponding author: Dr Helena A. Watson

Email: Helena.a.watson@kcl.ac.uk

Short Title: QUiPP v.2 asymp prediction

Keywords App, fetal fibronectin, cervical length, risk assessment, preterm

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.20401

Contribution:

What does this work add to what is already known?

This QUiPP app is the only decision-support tool for risk assessment of preterm birth in high-risk women. This extensive validation of the new algorithms which use quantitative fetal fibronectin, cervical length or both tests combined, provides crucial evidence for the popular app.

What are the clinical implications of this work?

Use of a such reliable tool in preterm surveillance clinics can assist in explanation of risk to women, facilitates complex management decisions and can help avoid unnecessary interventions.

Abstract

Objectives

Accurate mid-pregnancy prediction of spontaneous preterm birth (sPTB) is essential to ensure appropriate surveillance of high-risk women. Advancing the QUiPP prototype, QUiPP 2 aimed to provide individualised risk of delivery based on cervical length, quantitative fetal fibronectin (qfFN), or both tests combined, taking into account further risk factors, such as multiple pregnancy. Validation of the QUiPP 2 predictive models using a distinct dataset aims to confirm the accuracy and transportability of QUiPP overall and within certain clinically relevant timeframes.

Methods

This was a prospective secondary analysis of data from 13 UK preterm birth clinics. A total of 1803 (3878 visits) women were included in the training set and 904 women (1400 visits) in the validation set. Survival analysis was used to identify the significant predictors of sPTB and parametric structures for survival models were compared and the best selected. The estimated overall probability of delivery before six clinically important timepoints (30, 34, 37 weeks of pregnancy and within 1, 2 and 4 weeks of testing) was calculated for each woman and analysed as a predictive test for the actual occurrence of each event. This allowed receiver–operating characteristics (ROC) curves to be plotted, and areas under the curve (AUC) calculated. Calibration was performed to measure the agreement between expected and actual outcomes.

Results

All algorithms demonstrated high accuracy; AUCs between 0.75 and 0.90 for the use of qfFN and cervical length combined, 0.68 and 0.90 for qfFN and 0.71 and 0.87 for cervical length. The differences between the three algorithms were not statistically significant. Calibration confirmed no significant differences between expected and observed rates of sPTB within 4 weeks and slight over-estimation of risk with the use of cervical length measurement between 22-25^{+6} weeks' gestation.

Conclusion

QUiPP app version 2 is a highly accurate prediction tool for prematurity, based on a unique combination of biomarkers, symptoms and statistical algorithm. It can reliably be used in the context of discussing risk. Whilst further work is required to determine its role in identifying women

requiring prophylactic interventions, it is a reliable and convenient screening tool for planning follow-up or hospitalisation for high risk women.

Introduction

Premature birth (PTB) is a major determinant of lifelong health, and prevention of PTB of spontaneous onset is a key priority for service providers[1.] Cervical length (CL) and quantitative fetal fibronectin (qfFN) can assist in identification of those most at risk[2,3,] but their clinical utility may be limited by implementing treatments around a binary threshold which categorises women into one of high risk or low risk, without the appreciation that risk is a continuous variable[4.] Predictive modelling was incorporated into a decision-support tool in 2015 which provided an individualised risk of delivery within pre-specified timeframes for women based on use of CL and qfFN in combination. (https://quipp.org/)[5]. The prototype was installed on over 1,000 devices in 51 countries and our qualitative work suggested the app is a popular solution for communicating preterm birth risk with women. [6,7]

To enhance the accuracy and usability of the first version of the QUiPP app, we developed improved predictive models based on a larger dataset. For the convenience of clinicians in a both symptomatic and high-risk asymptomatic settings, the second version of the QUiPP app aimed to enable prediction based on qfFN and CL either alone or in combination, rather than requiring the results from both tests (i.e. six different algorithms). The inclusion of multiple pregnancies in the new models intended to enable prediction of spontaneous PTB (sPTB) in twin pregnancies using QUiPP rates. This is important because 50% of twins deliver preterm and multiple pregnancy rates are rising due to increased maternal age and assisted conception procedures [8.] The three algorithms for asymptomatic women were then validated on a separate dataset of high-risk asymptomatic women. In this paper we report the validation methods for development of the three asymptomatic QUiPP v.2 prediction models and their accuracy in estimating risk of sPTB within 4 weeks of testing and of testing between 22^{+0} and 25^{+6} weeks' gestation. These scenarios were chosen because they are likely to be useful contexts for managing and communicating risk for asymptomatic high-risk women. We are not recommending a particular cut-off or threshold to guide interventions in this paper because the app produces probabilities of delivery, which need to be interpreted in light of their clinical context. Development and validation of prediction models for use in the care of symptomatic women are described by this group's related paper ("*Development and validation of*

prediction models for QUIPP v.2, a tool for predicting preterm birth in symptomatic women" also submitted to this journal).

Methods

The QUiPP v.2 predictive models were generated by analysis using data collected from the prospective cohorts EQUIPP (REC number 10/H0806/68), INSIGHT (REC number 13/LO/0393) and POPPY studies between October 2010 and May 2017. Thirty five percent of the women (n=624) in the training dataset were also in the training set for the first version of the QUiPP app[5]. Since the models were destined for an updated version of a decision-making app, inclusion of all available data aimed to enhance predictive accuracy and there was no statistical or clinical reason to exclude the earlier data. The validation set included later participants from EQUIPP and INSIGHT studies between May 2017 and February 2019. This allowed us to test the models' performance on a new dataset and test transportability, given variations in the historical period but with a "plausible-related" sample (temporal validation)[9,10].

All women were asymptomatic and enrolled from 13 high-risk preterm surveillance clinics across the UK. Women were offered longitudinal qfFN testing using the quantitative rapid fFN analyzer (Hologic, Marlborough, MA, USA) every 2–4 weeks and/or transvaginal ultrasound CL measurement between 18^{+0} and 36^{+6} weeks' gestation. All women had at least one of the following risk factors for sPTB: previous preterm birth or prelabour rupture of membranes (PPROM) < 37 weeks, previous late miscarriage (16^{+0} -23^{+6} weeks), previous cervical surgery (e.g. LLETZ or cone biopsy) or twin pregnancy. Gestational age was confirmed with standard first trimester ultrasound scans. Each woman had cervical length measurement and/or qfFn measurement performed at the screening visit. The qfFN sample collection and transvaginal ultrasonic cervical length measurement were performed as previously described[11] . As per clinical protocols, the shortest cervical length measurement of three was used for this analysis. Women with a blood-stained swab or a history of vaginal douching or sexual intercourse within the last 24h were excluded from the study due to known interference with qfFN measurement. Participants' demographic characteristics, risk factors and obstetric history were entered into a secure online database (www.medscinet.net/PTBstudies). Women were managed as per unit protocols, with ultrasound-indicated cerclage offered if the shortest CL was measured as <25 mm prior to 24^{+0} week's gestation. Women with a cervix of <25

mm may also have been offered alternative treatments (vaginal progesterone or Arabin pessary) as part of relevant randomised-controlled trials (SuPPoRT, STOPPIT-2)[12,13.]

The primary studies were approved by South East London Research Ethics Committee (EQUIPP and POPPY) and City & East Research Ethics Committee (INSIGHT) and the local research ethics committees of all participating centres. Written informed consent was obtained from all participants.

Statistical methods

The QUIPP v.2 predictive models were created using CL and qffN data, CL alone and qfFN alone from the asymptomatic high-risk datasets. Statistical analysis was performed with Stata software (StataCorp, College Station, Texas).

Model generation

In developing the predictive models, there were two priorities:

1. We had a relatively small data set of high-quality information. We needed to make full use of all available data; every CL measurement or qfFN test carried out (according to Hologic's instructions), and for which the ultimate pregnancy outcome was known.
2. We aimed to produce a method of predicting prematurity that was simple enough to be programmed into a spreadsheet or a smartphone application; and accurate enough to aid clinical decision-making.

We made an early decision not to use logistic regression. This would have resulted in separate models for each endpoint; possibly producing inconsistent results. Instead, we opted to use survival analysis, so that a single, more powerful, model could estimate likelihood of delivery at any gestation. The modelling process is described in detail in the supplementary file: Statistical Methods for Creation of App Algorithms.

Model Validation

The validation set comprised high-risk asymptomatic women enrolled in the EQUIPP and INSIGHT studies up to February 2019 where outcomes had been collected since creation of the prediction models using the training set in May 2017. Multiple tests for individual women were included in the validation set because if a woman is tested on multiple occasions we wish to know the performance of every one of those test results. All the results included were taken at clinically appropriate times, so the results are all relevant to the validation process, and it is appropriate to include them. Selecting one episode per women would artificially reduce the dataset; and might introduce bias due to the selection process. Deliveries were classified as iatrogenic if labour was induced or a caesarean section was pre-labour, where membranes were intact. Induction of labour or CS following PPROM were treated as sPTB. For calculation of sPTB rates, if the birth was preterm (e.g. before 37 weeks) but iatrogenic, this data was counted as missing rather than excluded altogether since the data may be included within other outcomes (e.g. delivered within 7 days).

To estimate how well the QUiPP app discriminates between those who do, and those who do not, go on to have sPTB, the overall probability of delivery before six clinically important points in time (30, 34, 37 weeks of pregnancy and within 1, 2 and 4 weeks of testing) was calculated for each test event and analysed as a predictive test for the actual occurrence of each event. This allowed receiver–operating characteristics (ROC) curves to be plotted, and areas under the curve (AUC) calculated.

With preterm birth surveillance advisable for high-risk women[2, 1] we estimated that a four weekly interval was most likely to be feasible for most services and therefore we focused on achieving a high level of confidence in the algorithms ability to predict delivery within 4 weeks. To provide an accurate estimate of the uncertainty associated with all three algorithms' AUCs (for probability of sPTB within four weeks) bootstrapped confidence intervals were calculated. Bootstrapping allows for clustering by participant, in this case caused by repeated visits. To obtain an estimate of the bootstrap distribution 1,000 replications of the non-parametric resampling method were performed.

A sampling bias correction with acceleration was employed to allow for the fact that this data is not normally distributed.[14]

Calibration is another important method for measuring performance of predictive models and refers to the agreement between observed outcomes and predictions [15.] a calibration-in-the-large approach was used to compare the mean probability of delivery within 4 weeks with event rate and calibration-in-the-small to compare event rates with probabilities of <1%, 1-4.9%, 5-9.9% and > 10% delivery within 4 weeks. Calibration is essentially a check of accuracy. The aim of calibration-in-the-large is to check that the overall number of events actually observed is consistent with then number predicted. The aim of calibration-in-the-small is to check that groups of subjects with different predicted event rates do indeed have different actual event rates, as predicted

The performance of the app around 24 week's gestation is relevant to clinical decision-making regarding communicating risk and discharging women from preterm surveillance clinics, particularly at peri-viable gestations. For tests carried out between 22^{+0} and 25^{+6} weeks' gestation, the QUIPP v.2 performance in prediction of sPTB <30/34/37 weeks' gestation for each test group (CL alone, qfFN alone and both tests combined), was evaluated using ROC curves. When tests were repeated within this window, the first test result was used. Calibration-in-the-large and calibration-in-the-small was also performed to compare predicted with actual outcomes. For qfFN, CL and qfFN+CL taken between 22^{+0} and 25^{+6} weeks' gestation, probabilities of delivery before 34 and 37 weeks were compared with actual event rates and their binomial exact 95% confidence intervals.

Results

Model generation

Overall, 2303 asymptomatic women at high risk of sPTB from the EQUIPP, POPPY and INSIGHT prospective cohorts studies were selected from the trial database. Visits with incomplete qfFN/CL data (n=175), blood-stained swabs (n=85), invalid qfFNs (n=2), or sexual intercourse within 48 hours (n=233) were excluded from analysis. Triplets and one set of quads were excluded as prevalence of these higher order multiple pregnancies was too low to be useful but 150 sets of twins were included (with gestation of delivery of the first baby 1 used in the analysis) Following exclusions, CL and qfFN measurements from 1803 women at 3878 visits were analysed *(Figure 1).* This dataset including 288 spontaneous and 165 iatrogenic preterm deliveries. All measurements of qfFN and CL were taken between 18^{+0} and 36^{+6} weeks. The baseline characteristics for these women are described in *Table 1*.

Prediction models were created for use in three test groups (risk factors with CL alone, risk factors with qfFN alone and risk factors with CL and qfFN combined) from the training set. Parametric survival analysis with time-updated covariates was used to get the maximum information from the available data. For all three models the loglogistic 1 structure was the most accurate, using time of conception as time zero, even though women were observed only from time of test. For all three models, multiple pregnancy, previous sPTB and previous late miscarriage were required in the model. Other potential predictors (body mass index, smoking, ethnicity, and previous cervical surgery) were excluded as not significant in a multiple regression model. Smoothed hazard estimates using Cox regression suggest time-dependent risk factors for some variables including, qfFN, cervical surgery and late miscarriage. For example, *Figure 2* demonstrates the increasing risk of delivery with higher levels of qfFN and how even with low levels of qfFN, the hazard increases as you get beyond 200 days (28 weeks' gestation). This means that at earlier gestations risk is slightly over-estimated compared to a slight under-estimation of risk at late preterm gestations. Given that early preterm deliveries are most important and difficult to predict, this statistic anomaly actually incurs additional safety in the app, and a more complex statistical model for the app has not been sought.

Model Validation

For the validation set 1457 asymptomatic, high risk women with singleton or twin pregnancies were identified during the time period, with over 2211 visits. Exclusions were comparable with the training set and described in *Figure 1.* The complete validation dataset then included 1400 visits and 904 women (none of whom were included in the validation set for the first QUiPP algorithms).

The baseline demographics, risk factors and outcomes of the validation set were compared with the training set in *Table 1.* The validation set contains a significantly higher proportion of black women, and women with a previous sPTB or PPROM and significantly fewer Asian women and twin pregnancies. Whilst both datasets included women from across the UK, a greater contribution from INSIGHT recruits (70% of whom were recruited in an inner-city London hospital) in the later validation dataset is likely to account for these contrasts. This is an intended consequence of the temporal validation method which aims to assess the generalizability of the predictive models in different populations.

ROC curves for the qfFN, CL and qfFN/CL algorithms, based on all women in the dataset are shown (*Figure 3)*. All algorithms demonstrated good accuracy with areas under the curve (AUC) between 0.75 and 0.90 for the use of qfFN and CL combined, between 0.68 and 0.90 for qfFN and between 0.71 and 0.87 for CL. There were insufficient event rates for prediction within one and two weeks in this asymptomatic population to be useful. P values and confidence intervals for overall prediction have not been reported because no correction was made for repeated tests for most AUCs. Bootstrapped confidence intervals were calculated for delivery within 4 weeks of test were narrow supporting the models' reliability and no significant difference between the three algorithms (qfFN, CL, combined qfFN/CL) (*Table 2*).

Calibration for probability of delivery within 4 weeks included 1095 observations for qfFN alone, 988 observations for CL alone and 694 for combined use of qfFN and CL. Women were excluded if

spontaneous delivery before 4 weeks could not be determined because of iatrogenic delivery within this timeframe. *Tables 3 and 4* group women according to results on the QUiPP app. Calibration demonstrated no significant difference between the event rates and the predicted probabilities in any algorithm, confirming the QUiPP app's reliability at estimating risk of delivery within 4 weeks. (*Table 3)*. The qfFN algorithm in *Table 3* demonstrates the app's ability to segregate women according to true risk; a low (<1%) risk on the QUiPP app is associated with an event rate of only 0.5%, while a high risk of delivery within 4 weeks (>10%) probability is associated with a 26% risk of delivery within 4 weeks, compared to a no information rate of 2.9%. There are similar findings for the other biomarkers and other endpoints.

QUiPP v.2 predictive power at 22^{+0} to 25^{+6} weeks' gestation was good (AUC 0.763 for delivery <34 weeks and 0.746 for delivery <37 weeks using qfFN and CL) (*Figure* 4). For calibration at this gestation there were 1081 eligible observations for qfFN alone, 977 observations for CL alone and 689 observations for qfFN and CL. Event rates (sPTB < 34 and < 37 weeks) were similar to predicted rates of delivery using the qfFN algorithm for nearly all women. For the CL and CL/qfFN algorithms, the addition of CL data appeared to lead to statistically significant over-estimation of QUiPP risk, with predicted delivery rates above the upper confidence interval for actual delivery rates, particularly at higher risks (*Table* 4.)

Discussion

Created from data from over 1800 asymptomatic women, (three times the number used for creation of the prediction models used in first version of the QUiPP app) and validated on a significantly different population, this new version substantially increases the reliability and generalisability of QUiPP's prediction. For the first time, the new QUiPP app provides clinicians with three accurate methods for predicting preterm birth in high-risk women: using qfFN, CL or both tests together. Whilst there were trends towards improved prediction with both tests, these were not significant. The new version of the QUiPP app also incorporates calculation of risk in women with a twin pregnancy due to the inclusion of a large number of twin pregnancies (n=150) in the models' generation. Given uncertainties regarding the optimal management of twin pregnancies,[16,17,18] accurate identification of those most at risk of early preterm birth could help identify those most likely to benefit from intervention whilst reassuring the majority.

In this paper we have directly evaluated the performance of the QUiPP app, as used in the setting of a preterm surveillance clinic. This is timely since the NHS' new *Saving Babies' Lives Care Bundle* has made the provision of such clinics mandatory.[1] The QUiPP app may be a reliable tool in the establishment of preterm birth prevention pathways. We have confirmed excellent accuracy (AUCs > 0.87 for all three combinations of tests) for prediction of preterm birth within 4 weeks which is pertinent to clinicians planning serial follow up. Whilst the reported predictive performance are based on individual episodes, serial testing remains best practice[1,2,] will enhance the accuracy of tests further and is likely to be desirable to women with previous difficult experiences. The results from this study suggest that the app correctly predicts around half of visits with a < 1% risk of delivery within 4 weeks. At such a low level of risk, monthly follow-up may be acceptable, whereas an prompter appointment is likely to be appropriate for women with higher calculated risks. The models' accurate discrimination of women with a higher (e.g. > 10%) risk of delivery before 34 or 37 weeks at 22^{+0} to 25^{+6} weeks' gestation, means that QUiPP can reliably inform discussion of sPTB risk at the time women tend to be discharged from the specialist preterm clinic.

The app is a decision-making aid which we hope can make a valuable contributions to bespoke management plans, but we intentionally avoided providing a single cut-of. It is likely that a randomised-controlled trial design is required to address whether a specific QUiPP risk should prompt interventions such as cerclage or pessary. Furthermore, one of the key motivators for creating the app was that a single cut-off does not reflect the true utility of data across the risk range of variables such as cervical length and qfFN. Reducing the app output back to a binary threshold would under-estimate the importance of the clinical context and the woman's role in shared decision-making. In the absence of interventional trial data, we are not suggesting changing the approach for the majority of women at risk of preterm birth. However, given the reliability of the app's performance (Table 4), if at a peri-viable gestation a woman has a borderline CL but an overall risk of sPTB < 37 weeks < 10% (in line with background rates), conservative management seems justified.

Based on the results of our calibration, when QUiPP provides very low probabilities for preterm birth, event rates were consistently even lower providing a high level of reassurance to clinicians and women alike. In fact, there was a degree of over-estimation of risk, which was significant for the CL algorithms. It is possible that women with a short cervix did not deliver as soon as QUiPP predicted because these women received prophylactic interventions for sPTB, (e.g. cerclage, progesterone or Arabin pessary). However, this effect was significant for prediction at 22-25^{+6} weeks' when half of the women would not be eligible for prophylactic interventions, so treatment paradox may not fully explain this effect. Nevertheless, a degree of caution in the QUiPP app's estimations is likely to be protective and desirable, as avoiding the consequences of false negative results is more important than avoiding false positive results in sPTB risk assessment.

There were significant differences between our validation and training set which supports the QUiPP app's generalizability. In a vastly different population to either our training or validation set, cautious use of the QUiPP app is advised. In the future, as new insights into the pathophysiology of preterm parturition are realised and additional biomarkers are identified and large preterm birth

cohorts allow identification of new significant variables for the prediction models, there is potential to incorporate additional data into future iterations of QUiPP to refine prediction further.

The clinical impact of the QUiPP app for symptomatic women in terms of reducing inappropriate management of threatened preterm labour has been evaluated in a large randomized-controlled trial (EQUIPTT ISRCTN 17846337 due for publication late 2019). Similar trials are required for the asymptomatic population.

In conclusion, building on experiences and feedback from the first version, the second version of the QUiPP app provides enhanced usability and accuracy for risk assessment of high-risk women in preterm surveillance clinics. In accommodating a number of risk factors for sPTB, it is the only tool that provides women with a bespoke preterm birth risk, and can provide significant reassurance to women and clinicians, as well as health resource savings by preventing inappropriate admissions and treatments. Its rigorous validation and compliance with software application regulations (CE marked Sept 2017 MHRA Ref A015030) support its widespread clinical use within preterm birth surveillance protocols.

Conflicts of Interest

HAW, KK, NLH, PTS and RMT declare no conflict of interest. AHS is currently performing trials supported financially, paid to institute (Hologic, Biomedica), and donated samples (Partosure) to compare fFN, Actim Partus and Partosure. He is an advisor to NICE on preterm prediction tests.

Funding

The Insight study is supported by Tommy's Charity (no. 1060508); National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' (GSTT) and King's College London (KCL) National Health Service Foundation Trust and, Rose- trees Trust (Charity no. 298582). NLH is funded by a NIHR Doctoral Research Fellowship (DRF-2013-06-171) PTS is partly

funded by Tommy's Charity and by and by NIHR Collaboration for Leadership in Applied Health Research and Care, South London. HAW is funded by a GSTT and KCL Biomedical Research Council Clinical Training Fellowship Award. JC is supported by an NIHR/HEE CAT Clinical Doctoral Research Fellowship (CDRF-2013-04-026). The research was supported by the NIHR Biomedical Research Centre based at GSTT and KCL and/or the NIHR Clinical Research Facility. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. NHS England. NHS England Saving Babies' Lives Version Two: A care bundle for reducing perinatal mortality [Internet]. 2019. Available from: https://www.england.nhs.uk/publication/saving-babies-lives-version-two-a-care-bundle-for-reducing-perinatal-mortality/ [accessed 2019 Mar 25]
2. National Collaborating Centre for Women's and Children's Health (UK). Preterm Labour and Birth. National Institute for Health and Care Excellence: Clinical Guidelines. National Institute for Health and Care Excellence (UK): London, 2015.
3. Abbott DS, Hezelgrave NL, Seed PT, Norman JE, David AL, Bennett PR, Girling JC, Chandirimani M, Stock SJ, Carter J, Kurtzman J, Tribe RM, Shennan AH. Quantitative Fetal Fibronectin to Predict Preterm Birth in Asymptomatic Women at High Risk. Obstet Gynecol. 2015;125(5):1168.
4. Ridout A, Carter J, Shennan A. Clinical utility of quantitative fetal fibronectin in preterm labour. Vol. 123, BJOG. 2016. p. 1972.
5. Kuhrt K, Smout E, Hezelgrave N, Seed PT, Carter J, Shennan AH. Development and validation of a tool incorporating cervical length and quantitative fetal fibronectin to predict spontaneous preterm birth in asymptomatic high-risk women. Ultrasound Obstet Gynecol. 2016;47(1):104–9.
6. Carter J, Tribe RM, Shennan AH, Sandall J. Threatened preterm labour: Women's experiences of risk and care management: A qualitative study. Midwifery. 2018; 64:85-92.
7. Watson HA, Ridout A, Ross G, Shennan AH. Decision-making about preterm birth using the QUIPP app: A survey of women's experiences. Pregnancy Outcome Poster Abstracts BJOG An Int J Obstet Gynaecol. 2017;
8. Office of National Statistics. Birth characteristics in England and Wales - Office for National Statistics [Internet]. 2015.. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthcharacteristicsinenglandandwales/2015. [accessed 2019 Feb 11]
9. Altman DG, Royston P. What do we mean by validating a prognistic model? Stat Med. 2000;19:453–73.
10. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999; 130(6):515-24.
11. Bolt LA, Chandiramani M, De Greeff A, Seed PT, Kurtzman J, Shennan AH. The value of combined cervical length measurement and fetal fibronectin testing to predict spontaneous preterm birth in asymptomatic high-risk women. J Matern Neonatal Med. 2011;24(7):928–32.

12. Hezelgrave NL, Watson HA, Ridout A, Diab F, Seed PT, Chin-Smith E, Tribe RM, Shennan AH. Rationale and design of SuPPoRT: A multi-centre randomised controlled trial to compare three treatments: Cervical cerclage, cervical pessary and vaginal progesterone, for the prevention of preterm birth in women who develop a short cervix. BMC Pregnancy Childbirth. 2016;16(1):358.

13. Norman JE, Norrie J, Maclennan G, Cooper D, Whyte S, Burley SC, Smith JBE, Shennan A, Robson SC, Thornton S, Kilby MD, MArlow N, Stock SJ, Bennett PR, Denton J. Open randomised trial of the (Arabin) pessary to prevent preterm birth in twin pregnancy with health economics and acceptability: STOPPIT-2-a study protocol. BMJ Open. 2018; 8(12):e026430.

14. Carpenter J, Bithell J. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. Stat Med. 2000;

15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010. 21(1):128-38.

16. Norman JE, Mackenzie F, Owen P, Mactier H, Hanretty K, Cooper S, CAlder A, Mires G, Danielian P, Sturgiss S, MacLennan G, Tydeman G, Thornton S, MArton B, Thornton JG, Neilson JP, Norrie J. Progesterone for the prevention of preterm birth in twin pregnancy (STOPPIT): a randomised, double-blind, placebo-controlled study and meta-analysis. Lancet. 2009; 373(9680):2034-40.

17. Nicolaides KH, Syngelaki A, Poon LC, de Paco Matallana C, Plasencia W, Molina FS, Picciarelli G, Tul N, Celik E, Lau TK, Conturso R. Cervical pessary placement for prevention of preterm birth in unselected twin pregnancies: a randomized controlled trial. Am J Obstet Gynecol [Internet]. 2015;214(1):3.e1-3.e9. A

18. Berghella V, Odibo AO, To MS, Rust OA, Althuisius SM. Cerclage for short cervix on ultrasonography: meta-analysis of trials using individual patient-level data. Obstet Gynecol. 2005;106(1):181–9.

Figure legends

Figure 1: Eligibility flowchart for QUiPP 2 asymptomatic predictive model datasets

Figure 2:Cox regression smoothed hazard estimates be gestation for different categories of quantitative fFN

Figure 3: ROC curves demonstrating the overall prediction values for quantitative fetal fibronectin, cervical length and both tests combined using QUiPP app. ROC curves presented for delivery < 30, 34 and 37 weeks and within one week from test.

Figure 4: ROC curves for QUiPP prediction using quantitative fetal fibronectin, cervical length and both tests combined at 22+0-25+6 week's gestation

Table 1: Baseline Characteristics of asymptomatic women providing data for model generation and validation of the Quipp app algorithm datasets

Characteristic	**Training set** (n=1803)	**Validation set** (n=904 women)	**Comparison** (95% CI)
Demographics			*Difference between means*
Age (y)	32.9 ± 5.20	32.8 ± 5.13	0.18 (-0.23 to 0.59)
BMI (kg/m^2)	25.6 ± 4.96	25.4 ± 5.45	0.38 (-0.04 to 0.81)
Ethnicity			*Relative Risk*
• White	985 (54.7)	606 (67.0)	0.51 (0.37 to 0.71)*
• Black	59 (3.3)	75 (8.3)	1.90 (1.62 to 2.23)*
• Asian	572 (31.8)	145 (16.0)	1.39 (1.08 to 1.78)*
• Other	185 (10.3)	78 (8.6)	
PTB Risk Factors			
Previous PTB	563 (30.9)	417 (46.1)	0.67 (0.61 to 0.74)*
Previous PPROM	252 (13.8)	267 (29.5)	0.47 (0.40 to 0.55)*
Previous Late Miscarriage	388 (21.3)	198 (21.9)	0.97 (0.84 to 1.13)
Previous Cervical Surgery	702 (38.6)	397 (43.9)	0.88 (0.80 to 0.96)*
Twin pregnancy	136 (7.5)	17 (1.9)	3.97 (2.42 to 6.54)*
Number of visits per woman (mean)	2.17	1.53	0.64 (0.54 to 0.73)*
Pregnancy Outcomes			
Gestation at delivery (days/weeks)	266 /38^{+0}	267/38^{+1}	-1.28 (-3.28 to 0.73)
sPTB <37 weeks	231 (13.8)	117 (13.8)	0.99 (0.81 to 1.22)
Induced labour	380 (21.1)	181 (20.4)	1.08 (0.94 to 1.25)
Pre-labour LSCS	349 (19.4)	144 (16.2)	0.98 (0.93 to 1.03)

Data given as mean or n (%) and standard deviation

*Denotes significant difference between training and validation sets

Table 2: QUiPP app predictive accuracy of delivery within 4 weeks of testing

Prediction timeframe	Area under the curve (AUC) (bootstrapped confidence intervals)		
	fFN alone (m-1095)	CL alone (n=988)	fFN and CL (n-694)
<4 weeks	0.866 (0.784-0.927)	0.865 (0.720-0.919)	0.888 (0.728-0.953)

Table 3: QUiPP app calibration results for delivery within 4 weeks

fFN alone (n=1095)			
Predicted probability (calculated by QUiPP)	Predicted event rate *	Actual event rate x/n (%)	95% CI of actual event rate
<1% (561)	0.6%	0.5% (3)	0.1 to 1.6
1 to 4.9% (424)	2.0%	2.4% (10)	1.1 to 4.3
5 to 9.9% (57)	6.9%	8.8% (5)	2.9 to 19.3
>10% (53)	23.5%	26.4% (14)	15.3 to 40.3
All women (calibration in the large)	2.57%	2.92%	2.0 to 41.0
CL alone (n=988)			
Predicted probability (calculated by QUiPP)	Predicted event rate *	Actual event rate x/n (%)	95% CI of actual event rate
<1% (499)	0.5%	0.2% (1)	0.0 to 1.1
1 to 4.9% (354)	2.1%	1.9% (7)	0.8 to 4.0
5 to 9.9% (55)	7.0%	5.5% (3)	1.1 to 15.1
>10% (80)	24.6%	13.8% (11)	7.1 to 23.2
All women (calibration in the large)	3.4%	2.2%	1.4 to 3.4
fFN and CL (n=694)			
Predicted probability (calculated by QUiPP)	Predicted event rate *	Actual event rate x/n (%)	95% CI of actual event rate
<1% (383)	0.4%	0.3% (1)	0.0 to 1.4
1 to 4.9% (204)	2.1%	1.5% (3)	0.3 to 4.2
5 to 9.9% (41)	6.7&	0.0% (0)	0.0 to 8.6
>10% (66)	24.1%	18.2% (12)	9.8 to 29.6
All women (calibration in the large)	3.6%	2.3%	1.3 to 3.7

* The predicted event rate is the average of the predicted probabilities, as calculated by the QUiPP app, for all the women in the group.

Table 4: QUiPP app calibration results for tests performed between 22-25+6

Predicted probability delivery < **34** weeks (calculated by QUiPP)	**fFN alone (n=1081)**		
	Predicted event rate *	Actual event rate x/n (%)	95% CI of actual event rate
1-4.9%	3.5%	2.2% (9)	1.0 to 4.1
5-9.9%	6.9%	7.3% (33)	5.0 to 10.0
>10%	25.5%	20.2% (44)	15.1 to 26.1
Overall (calibration in the large)	9.3%	8.0% (86)	6.4 to 9.7
Predicted probability delivery < **34** weeks (calculated by QUiPP)	**CL alone (n=977)**		
	Predicted event rate	Actual event rate x/n (%)	95% CI of actual event rate
1-4.9%	3.5%	2.3%(5)	0.7 to 5.3
5-9.9%	7.3%	2.9% (10)	1.4 to 5.3
>10%	24.1%	14.2% (59)	11 to 18
Overall (calibration in the large)	13.6%	7.6% (74)	6.0 to 9.4
Predicted probability delivery < **34** weeks (calculated by QUiPP)	**fFN and CL (n=689)**		
	Predicted event rate	Actual event rate x/n (%)	95% CI of actual event rate
1-4.9%	3.8%	0.0% (0)	0.0 to 4.5
5-9.9%	7.5%	4.4% (11)	2.2 to 7.7
>10%	24.7%	13.7% (49)	10.3 to 17.7
Overall (calibration in the large)	16.0%	8.7%(60)	6.7 to 11.1
Predicted probability delivery < **37** weeks (calculated by QUiPP)	**fFN alone (n=1020)**		
	Predicted event rate *	Actual event rate x/n (%)	95% CI of actual event rate
1-4.9%	4.5%	3.0% (2)	0.4 to 10.5
5-9.9%	7.2%	10.4% (41)	7.6 to 13.8
>10%	21.3%	23.3% (130)	19.8 to 27.0
Overall (calibration in the large)	14.8%	17.0% (173)	14.7 to 19.4
Predicted probability delivery < **37** weeks (calculated by QUiPP)	**CL alone (n=935)**		
	Predicted event rate	Actual event rate x/n (%)	95% CI of actual event rate
1-4.9%	3.4%	2.0% (1)	0.0 to 10.5
5-9.9%	7.6%	6.3% (13)	3.4 to 10.5
>10%	26.0%	21.3% (144)	118.3 to 24.6
Overall (calibration in the large)	20.7%	16.9% (158)	14.6 to 19.5
Predicted probability delivery < **37** weeks (calculated by QUiPP)	**fFN and CL (n=656)**		
	Predicted event rate	Actual event rate x/n (%)	95% CI of actual event rate
1-4.9%	4.3%	0.0% (0)	0.0 to 30.1
5-9.9%	8.0%	3.0% (3)	0.6 to 8.5
>10%	27.3%	21.1% (115)	17.7 to 24.5
Overall (calibration in the large)	24.0%	18.0% (118)	15.1 to 21.1

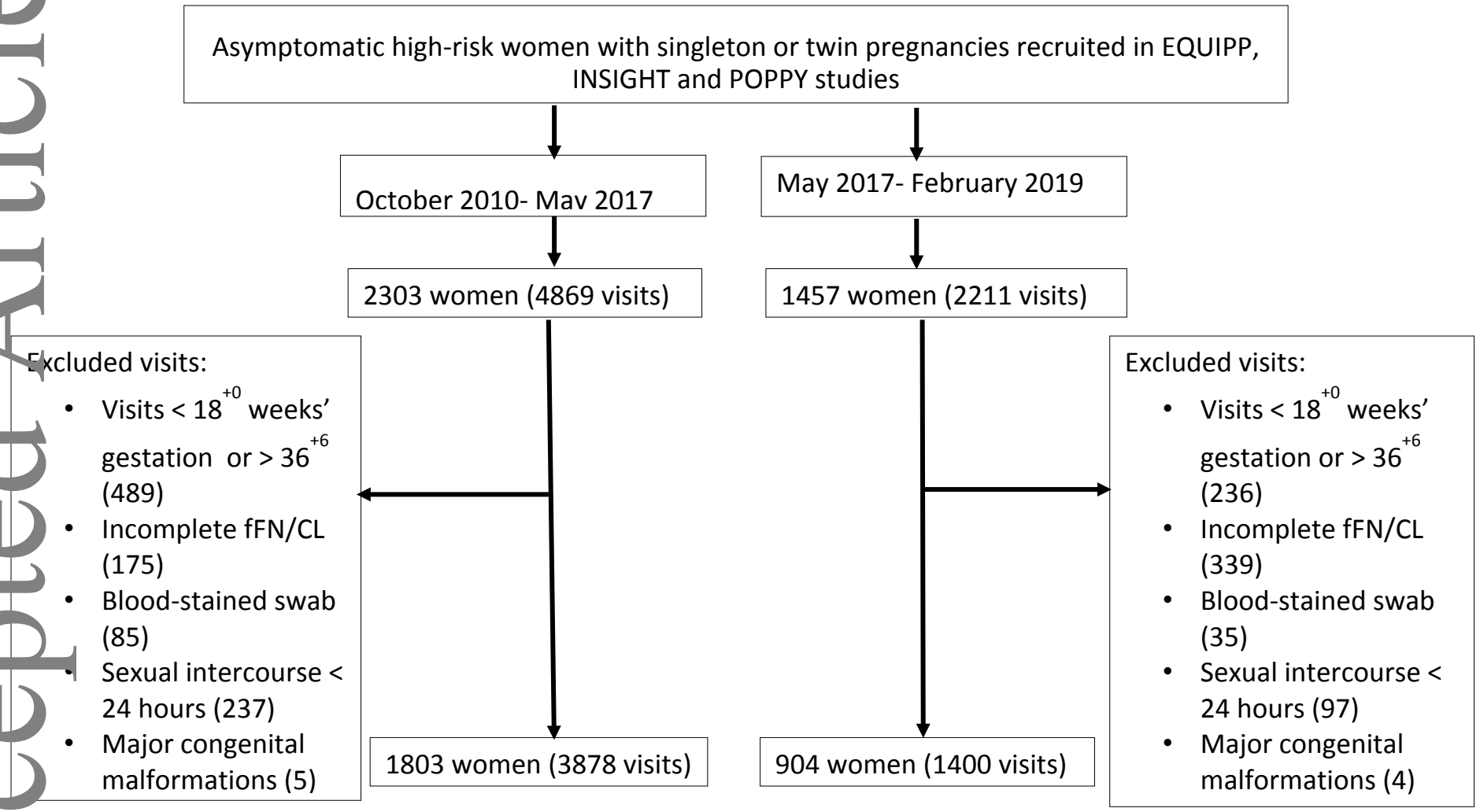
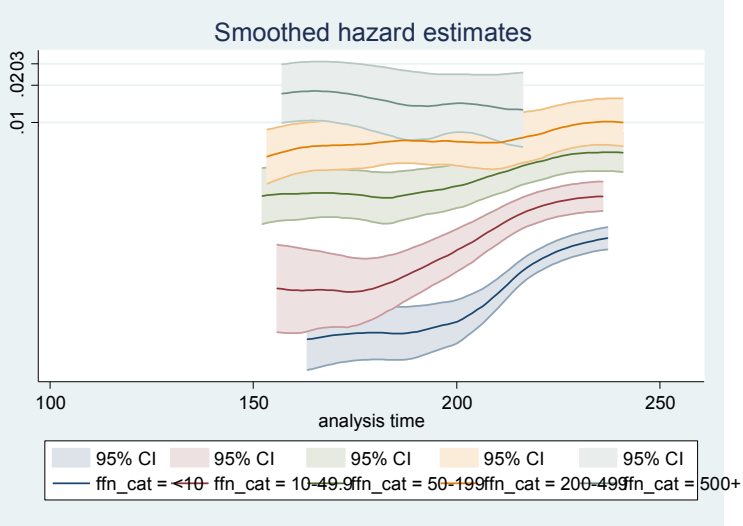
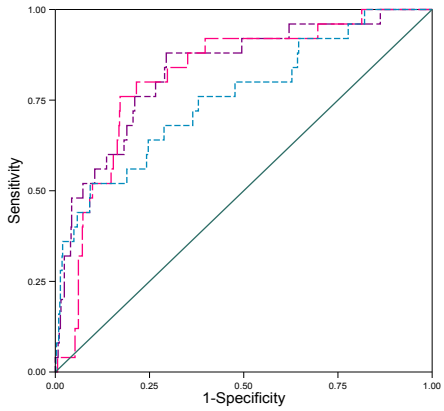


Figure 1: Eligibility flowchart for QUiPP 2 asymptomatic predictive model datasets

Figure 2:Cox regression smoothed hazard estimates be gestation for different categories of quantitative fFN

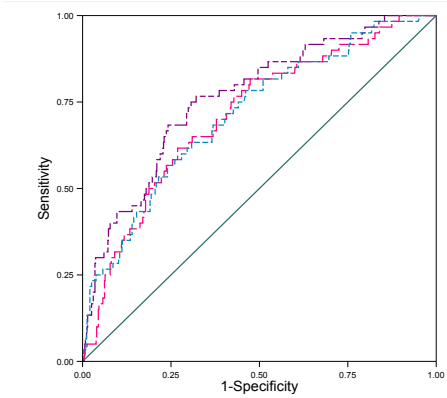


ROC for overall prediction of delivery < 30 weeks



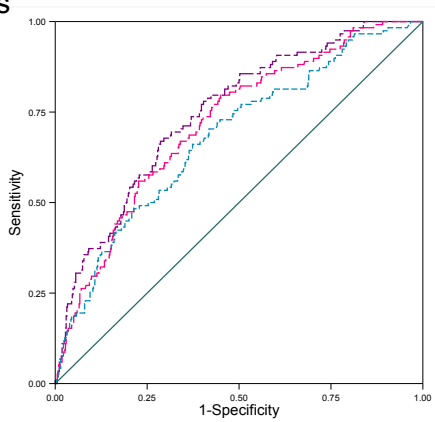
AUC qfFN 0.756. CL 0.817. CL/fFN

ROC for overall prediction of delivery < 34 weeks



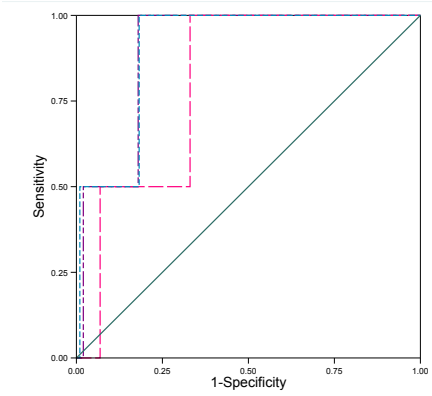
AUC qfFN 0.714, CL 0.713, CL/fFN 0.763

ROC for overall prediction of delivery < 37 weeks



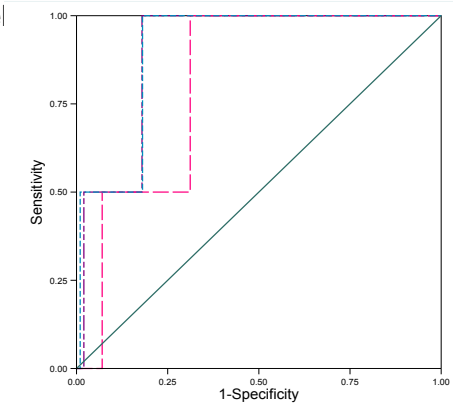
AUC qfFN 0.680, CL 0.717, CL/fFN 0.746

ROC for overall prediction of delivery within 1 week



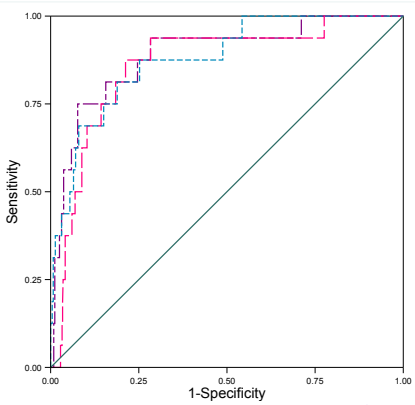
AUC qfFN 0.904, CL 0.800, CL/fFN 0.900

ROC for overall prediction of delivery within 2 wee



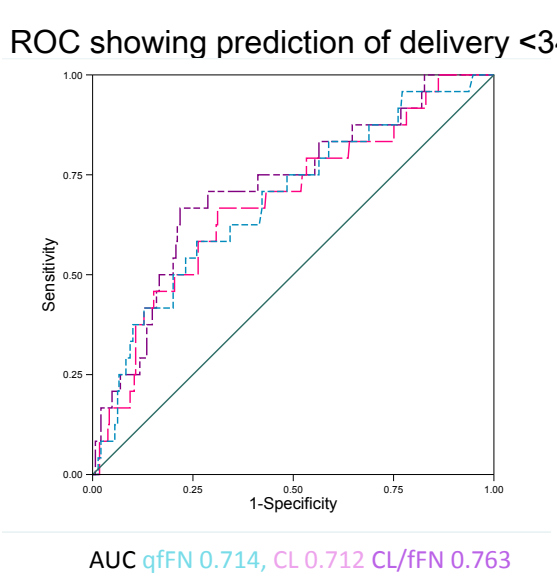
AUC qfFN 0.90, CL 0.809, CL/fFN 0.900

ROC for overall prediction of delivery within 4 .



AUC qfFN 0.866, CL 0.865, CL/fFN 0.888

Figure 3: ROC curves demonstrating the overall prediction values for quantitative fetal fibronectin, cervical length and both tests combined using QUiPP app. ROC curves presented for delivery < 30, 34 and 37 weeks and within one week from test.



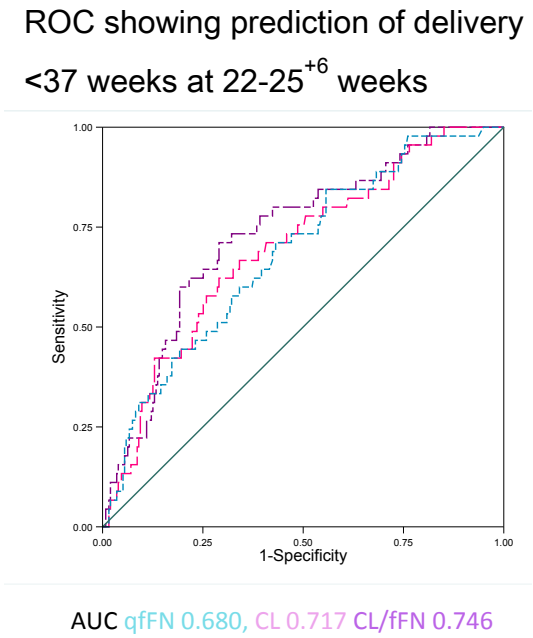


Figure 4: ROC curves for QUiPP prediction using quantitative fetal fibronectin, cervical length and both tests combined at 22+0-25+6 week's gestation