Archives of Gynecology and Obstetrics (2022) 306:571–575
https://doi.org/10.1007/s00404-022-06418-2

CORRESPONDENCE

The web-based application "QUiPP v.2" for the prediction of preterm birth in symptomatic women is not yet ready for worldwide clinical use: ten reflections on development, validation and use

Isabelle Dehaene[1] · Johan Steen[2,3] · Gilles Vandewiele[4] · Kristien Roelens[1] · Johan Decruyenaere[3]

Received: 2 June 2021 / Accepted: 24 January 2022 / Published online: 2 February 2022

Abstract

Purpose In this correspondence, we highlight general and domain-specific caveats in the development and validation of prediction models.

Methods Development and use of the "QUiPP" application, a tool for preterm birth prediction which is supported by the United Kingdom National Health Service, is scrutinised and commented on.

Results We highlight and elaborate ten points which may be perceived to be unclear or potentially misleading.

Conclusion While the QUiPP application has high potential, it lacks transparency (on certain aspects related to model development) and proper validation. This precludes transportability to settings with other treatment policies and to other countries where the app has been made publicly available.

Keywords Preterm birth prediction · Fibronectin · Cervical length · eHealth · QUiPP

The QUiPP (Quantitative Instrument for the Prediction of Preterm birth) application (app) was developed as a tool to predict the individual probability of spontaneous preterm birth (sPTB) in symptomatic and asymptomatic women based on risk factors for preterm birth, gestational age at testing, and cervical length measurement and/or quantitative foetal fibronectin (qfFN). The app is supported by the United Kingdom (UK) National Health Service (NHS), where it was developed, and is also used in other countries since it was launched as a mobile app in 2017. In the UK, it has even been used during the SARS-CoV-2 pandemic to help decrease unnecessary admissions and transfers [1, 2].

In this correspondence, we will focus on the QUiPP app prediction of preterm birth risk in symptomatic women. We summarise how the app was developed and validated, and comment on both these processes and on the practical use of the app.

Development and validation of the app

The first version of the QUiPP app was developed based on a secondary analysis of the EQUIPP (Evaluation of fetal fibronectin with a Quantitative Instrument for the Prediction of Preterm birth) study dataset [3]. This study included 382 symptomatic women with singleton pregnancies who underwent qfFN testing between 22^{+0} and 35^{+6} weeks. Women were excluded (24.2%, 122 of 504 eligible women) when no fFN swab was available ($n = 24$, 4.8%) and in case of multiple pregnancy, congenital malformations, or incomplete outcome data ($n = 35$, 6.9%). The remaining dataset was split in a training set ($n = 190$) and test set ($n = 192$ subsequently admitted women) that was used for temporal external validation. All women with a positive qfFN test were managed as per unit protocols (antenatal corticosteroids, tocolysis, bed rest). Women were considered to be at risk of sPTB until birth or 37 weeks' gestation. The rate of sPTB in both training and test set was 13%. There were in total 3% iatrogenic

✉ Isabelle Dehaene
isabelle.dehaene@ugent.be

[1] Obstetrics and Gynaecology, Ghent University Hospital, Corneel Heymanslaan 10, 9000 Ghent, Belgium

[2] Department of Internal Medicine and Paediatrics, Renal Division, Ghent University, Ghent, Belgium

[3] Department of Intensive Care Medicine, Ghent University Hospital, Ghent, Belgium

[4] IDLab, Ghent University-IMEC, Ghent, Belgium

PTBs and 84% term births, which were considered as non-events and, therefore, censored at 37 weeks.

Multiple parametric models with different survival functions and predictor sets were fitted on the training set. The final selected model, having the lowest value of the Akaike and Bayesian information criteria, was a log-normal survival model with terms for qfFN and previous sPTB or preterm prelabour rupture of membranes (PPROM). This final model was used to calculate the probability of delivery before 30, 34, and 37 weeks' gestation and the probability of delivery within 2 or 4 weeks from qfFN testing. Considering a probability of delivery of more than 10% as a positive test, the resulting markers were characterised by having low positive and high negative predictive value.

The QUiPP app was updated in 2019 [4]. Further development and validation of the algorithms had been done, introducing additional risk factors as predictors and enabling risk calculation using either qfFN, cervical length (CL) measurement, or both. Data from four different prospective cohort studies were used to update the app. For the goal of model development, women were considered eligible if they had symptoms of threatened PTB between 23^{+0} and 34^{+6} weeks' gestation. They were excluded ($n=$ unknown) when labour was established, membranes ruptured, or when there was antenatal haemorrhage. Additional exclusion criteria ($n=222$, 12.6%) were: invalid visits/qfFN test results ($n=18$, 1.0%), sexual intercourse < 24 h prior to testing, major congenital abnormality, incomplete outcome data ($n=42$, 2.4%), indication for iatrogenic PTB at presentation ($n=97$, 5.5%), and higher order multiple pregnancies. In total, there were 1032 participants included in the training set, including the initial EQUIPP cohort. For validation of the models (temporal external validation), 506 women from one of the cohorts were used (test set). Three prediction algorithms were developed, using Cox proportional hazards regression: one using only qfFN (n training = 1032, n test = 506), one using only CL measurement (n training = 204, n test = 132), and one using both tests (n training = 204, n test = 128). In the cohort with both qfFN and CL test results, only previous cervical surgery was found to provide added predictive power to qfFN, CL, and gestational age at test. However, to maintain consistency with the corresponding algorithm for risk prediction in asymptomatic women, the decision was made to use the same variables: "previous cervical surgery?" (yes/no), "previous spontaneous preterm birth ≤ 36^{+6}?" (yes/no), "previous PPROM?" (yes/no), "number of foetuses?" (1 or 2), "gestation at test?" (18–36 weeks), and "shortest cervical length" (mm) or "qfFN result" (ng/ml). The test was considered positive when the predicted risk was higher than 5% (as opposed to 10% in the first version).

Reflection on development, validation and use of the app

We identified 10 ambiguities in the manuscripts on the development of the app [3, 4] and on the website on which the suggested use of the QUiPP v.2 app is formulated (https://quipp.org/about.html).

(1) In 2015, 4 years before the publication of the second version of the QUiPP app, Collins G et al. published the TRIPOD statement: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis [5]. This statement aims to improve the transparency of the reporting of a prediction model study. Application of the checklist points out that there are some shortcomings in the reporting of the study by Carter et al. (Online Resource 1, Checklist TRIPOD) [4]. Three items will be further discussed in this correspondence: not all outcome variables and predictors were explicitly defined (point 2), unclear and non-intuitive (even counter-intuitive) handling of the predictors (point 3), and unclear description of the participants flow (point 4).

(2) The definition of previous PPROM, a risk factor used as a predictor for making individualised risk predictions, is not described in the app, nor is it easily found in the article [4]. Goodfellow et al. contacted the authors and the following definition was provided: previous PPROM is membrane rupture before 37 weeks followed by birth after 37 weeks [6].

(3) Close inspection of the algorithm formulas reveals that risk factors (other than the test results and gestation at test) do not have an individual coefficient in the formula but share a mutual coefficient. In other words, no distinction is made in the individual weight of each risk factor nor in the number of risk factors present, as the mutual coefficient is multiplied by an indicator of the presence of at least one risk factor. This is counter-intuitive for clinicians who use the app.

(4) It is not entirely clear how iatrogenic PTBs were handled in the study of Carter et al. [4]. According to the flow chart that displays participants after exclusions and split between training and validations sets, 97 women were excluded because they experienced an iatrogenic PTB (97/1760, 5.5%). In the main text, however, it is stated that iatrogenic PTBs were not excluded, but were instead treated as non-events, similar to term births, and were therefore censored at 37 weeks. It is unclear whether and in which way the included iatrogenic PTBs ($n=$ unavailable) differed from those that were excluded. It is also difficult for the reader to discern which predictions the authors

had in mind. Was their goal to predict the risk of sPTB under the hypothetical assumption that no women experienced an iatrogenic PTB? Or was their goal to predict the risk of sPTB under the realistic assumption that iatrogenic delivery may become indicated at some point after study inclusion? Our guess is that the excluded iatrogenic PTBs concerned deliveries with clear medical indications for planned PTB at study inclusion, while those that were included but censored at 37 weeks' gestation concerned deliveries for which a medical indication for planned PTB that occurred only after study inclusion and testing. From a clinical perspective, this would make sense because we believe the authors (i) did not wish to make individualised risk predictions in women for whom iatrogenic delivery is already planned (i.e. a plausible motivation for why certain iatrogenic PTBs were excluded) and (ii) did not wish to exclude the possibility of iatrogenic delivery (i.e. a plausible motivation for why certain iatrogenic PTBs were artificially censored at 37 weeks instead of censored at their respective time-of-delivery). These analytic choices were, however, poorly documented and even more poorly motivated. We, therefore, believe the authors missed the opportunity of providing a clear description of a well-defined risk [i.e. risk without (hypothetical) elimination of iatrogenic PTB as a competing event] in a well-defined target populations (i.e. patients at risk of sPTB, but without clear medical indications for iatrogenic delivery at the time of testing) [7, 8].

(5) In the background of the application, three different prediction models are being used (qfFN alone, CL alone, CL + qfFN). Predictive accuracy is different for each model, for each outcome, and in different settings. However, this is not transparent for the clinician for who the app is intended. For example, the area under the receiver operating curve (AUROC) for the prediction of birth within 7 days after testing was 69.8% for the CL model versus 87.5% for the CL + qfFN model. The AUROC for the prediction of birth before 30 weeks' gestation was 84.8% for the CL model versus 95.3% for the CL + qfFN model. Predictive accuracy parameters, other than the AUROC, also depend on the threshold used to identify patients at increased risk and accordingly classify as 'positives'. The higher the threshold, the lower the sensitivity and the higher the specificity. The NHS handles a "treat all" policy at a gestational age less than 30 weeks. For this reason, the threshold considered clinically relevant is low (5%), which is reflected in the high negative predictive value and low positive predictive value of all models and for all outcomes. According to the Belgian national guideline, for example, treatment depends on CL measurement and qualitative fFN or phosphorylated insulin-like growth factor binding protein 1 (phIGFBP-1) testing, which already results in less admissions and treatment initiations below 30 weeks [9]. The added value of the QUiPP app in this and other settings is to be explored.

(6) While it is mentioned on the website that the symptomatic algorithm is suitable for women between 23^{+0} and 34^{+6} weeks' gestation, the app also provides risk predictions for symptomatic women at 18–23 and 35–36 weeks. This is model-based extrapolation beyond the range of the training set.

(7) When the outcome was unknown or test results were invalid or missing, patients were excluded. It is unclear why the outcome was missing in some study participants. Whenever the missingness pattern is non-random, exclusion of patients due to missing outcomes (e.g. in complete case analyses) may lead to selection bias and miscalibration with respect to the intended population. This issue may also occur when patients are initially included but lost to follow-up during the course of the study and their (unobserved) outcomes are therefore censored at loss to follow-up. Importantly, miscalibration may go unnoticed when also the validation set is plagued by non-random missingness (especially when this is governed by the same underlying factors). Statistical techniques, such as inverse probability (of censoring) weighting, may be used to reduce selection bias due to informative missingness or censoring, but only to the extent that the censoring mechanism can be explained by measured covariates. Exclusion of women for whom no (valid) fFN test result was available, may likewise have led to selection bias. One could, however, argue that the risk of significant miscalibration in the development of the QUiPP app due to exclusion of patients with unknown or invalid fFN or outcome is probably low, since the number of such exclusions is relatively low ($n = 60/1760$, 3.4%).

(8) In the datasets used to develop the QUiPP app, treatment of the women was according to per unit protocols. Accordingly, risk predictions produced by the app implicitly capture potential treatment effects. In real life, the app is or will probably be used at the time patients present with symptoms, without having received treatment yet, since the tool is developed to support clinical decisions on treatment and interventions. The risk of the untreated patient to deliver within 7 days after presentation, however, might be higher than the risk of the same patient in case she had received standard-of-care treatment after presentation, which may e.g. include tocolysis. This may lead to a phenomenon called the "treatment paradox": a strong

predictor of the outcome triggers an effective treatment that may, in turn, prevent or alter (e.g. delay) the outcome, such that the association between that predictor and the outcome is diluted or even reversed compared to the hypothetical scenario in which no treatment had been given [10, 11]. Although this issue is touched upon by Carter et al., the authors do not fully discuss its potential implications and seem to ignore or to be unaware of the fact that this may be problematic when using the app to guide decision making [4].

(9) Another limitation is the lack of any measure that quantifies the degree of uncertainty around the risk predictions. Risk predictions are estimates provided by statistical models which are characterised by uncertainty. This uncertainty also provides valuable information for the clinician.

(10) And finally (10), the app was made publicly available before external validation was performed. In 2021, a trial was published which explored if the QUiPP app could prevent unnecessary management. The cohort of this trial was used to perform external validation of the app [12]. Visits and/or women were excluded when the cervical length was measured in addition to fFN (the reason for this is not stated), when there was no documented onset of labour and gestation of delivery, and in case of iatrogenic birth within the specified time period. The latter shows that our guess on why and when iatrogenic births were excluded (point 4), was not correct or that the researchers not consistently followed their proper patient flow. Excluding patients based on events that occur after the prediction landmark without statistical correction for this non-random selection may introduce bias in the assessment of the predictive performance of the models. The authors conclude that the ROC curves support the use of the tool to triage threatening preterm birth; however, they show that without the tool, the number of unnecessary treatments is not higher and there seems to be no difference in prognostic accuracy between using the QUiPP app and the sole use of fFN.

Moreover, external validation studies in other countries with different treatment policies are mandatory to justify its use outside the UK.

Conclusion

The QUiPP app has several strengths: it is user friendly, aimed to be used on the spot, and requires only a limited number of easy to collect variables. The developers aim to better estimate and individualise risk predictions, which is to be applauded, as well as publicly sharing their algorithm code and patient data, thereby fostering open research. Unfortunately, to date, the app lacks proper validation, and hence generalisability and transportability. Its widespread availability and use should, therefore, be discouraged until appropriate validation of all models, reported following the TRIPOD guidelines, is done in other settings and countries, or alternatively according to well specified user criteria. If the above points were to be addressed, a reliable, transparent, and helpful app could be the result.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00404-022-06418-2.

Author contributions ID: Conceptualisation, Writing—Original Draft, Writing—Review and Editing. JS: Writing—Original Draft, Writing—Review and Editing. GV: Writing—Original Draft, Writing—Review and Editing. KR: Supervision. JD: Supervision.

Funding Isabelle Dehaene is funded by a scholarship of FWO (1700520N). The funding body played no role in the creation of this manuscript.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. BAPM.org: QuiPP App Toolkit. https://www.bapm.org/pages/187-quipp-app/. Accessed on 28 June 2020
2. Carlisle N, Watson H, Shennan A (2021) Development and rapid rollout of The QUiPP App Toolkit for women who arrive in threatened preterm labour. BMJ Open Quality 10:e001272. https://doi.org/10.1136/bmjoq-2020-001272
3. Kuhrt K, Hezelgrave N, Foster C, Seed P, Shennan A (2016) Development and validation of a tool incorporating quantitative fetal fibronectin to predict spontaneous preterm birth in symptomatic women. Ultrasound Obstet Gynecol 47(2):210–216
4. Carter J, Seed P, Watson H, David A, Sandall J, Shennan A, Tribe R (2020) Development and validation of prediction models for the QUiPP app vol 2: a tool for predicting preterm birth in women with symptoms of threatened preterm labour. Ultrasound Obstet Gynecol 55(3):357–367
5. Collins G, Reitsma J, Atlman D, Moons K (2015) Transparant reporting of a multivariable prediction model for individual

prognosis and diagnosis (TRIPOD): the TRIPOD statement. BMC Med 13:1. https://doi.org/10.1186/s12916-014-0241-z

6. Goodfellow L, Care A, Sharp A, Ivandic J, Poljak B, Roberts D, Alfirevic Z (2019) Effect of QUiPP prediction algorithm on treatment decisions in women with a previous preterm birth: a prospective cohort study. BJOG 126(13):1569–1575
7. Young J, Stensrud M, Tchetgen E, Hernán M (2020) A causal framework for classical statistical estimands in failure-time settings with competing events. Stat Med 39(8):1199–1236. https://doi.org/10.1002/sim.8471
8. van Geloven N, Swanson S, Ramspek C, Luijken K, van Diepen M, Morris T, Groenwold R, van Houwelingen J, Putter H, Le Cessie S (2020) Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol 35:619–630. https://doi.org/10.1007/s10654-020-00636-1
9. Roelens K, Roberfroid D, Ahmadzai N, Ansari M, Singh K, Gaudet L, Alexander S, Cools F, de Thysebaert B, Emonts P, Faron G, Gyselaers W, Kirkpatrick C, Lewi L, Logghe H, Niset A, Rigo V, Tency I, Van Overmeire B,Verleye L (2014) Prevention of preterm birth in women at risk: selected topics. Good Clinical Practice (GCP) Brussels:Belgian Health Care Knowledge Centre (KCE). KCE reports 228. D/2014/10.273/63
10. Cheong-See F, Allotey J, Marlin N, Mol B, Schuit E, ter Riet G, Riley R, Moons K, Khan K, Thangaratinam S (2016) Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. BJOG 123(7):1060–1064
11. Schulam P, Saria S (2017) Reliable decision support using counterfactual models. Adv Neural Inf Process Syst 30:1698–1709
12. Watson H, Carlisle N, Seed P, Carter J, Huhrt K, Tribe R, Shennan A (2021) Evaluating the use of the QUiPP app and its impact on the management of threatened preterm labour: a cluster randomised trial. PLoS Med 18(7):e1003689