Prognostication for prelabor rupture of membranes and the time of delivery in nationwide insured women: development, validation, and deployment

Herdiantri Sufriyana, MD, MSc;[a,b] Yu-Wei Wu, PhD;[a,c] Emily Chia-Yu Su, PhD;[a,c,d,*]

[a] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan.

[b] Department of Medical Physiology, Faculty of Medicine, Universitas Nahdlatul Ulama Surabaya, 57 Raya Jemursari Road, Surabaya 60237, Indonesia.

[c] Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan.

[d] Research Center for Artificial Intelligence in Medicine, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan.

* Corresponding author at: Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. Phone: +886-2-66382736 ext. 1515.

Email addresses: herdiantrisufriyana@unusa.ac.id (HS); yuwei.wu@tmu.edu.tw (YWW); and emilysu@tmu.edu.tw (ECYS)

Title: 143 characters (with spaces)

Key points: 98 words

Abstract: 322 words

Main text: 3,000 words

Key Points

Question: Can we use medical histories of diagnosis and procedure in electronic health records to predict prelabor rupture of membranes and the time of delivery before the day in nationwide insured women?

Findings: In this prognostic study applying retrospective cohort paradigm, a significant predictive performance was achieved and validated. The area under receiver operating characteristics curve was 0.73 with the estimation errors of ± 2.2 and 2.6 weeks for the time of delivery.

Meaning: Preliminary prediction can be conducted in a wide population of insured women to predict prelabor rupture of membranes and estimate the time of delivery.

Abstract

Importance: Prognostic predictions of prelabor rupture of membranes lack proper sample sizes and external validation.

Objective: To develop, validate, and deploy statistical and/or machine learning prediction models using medical histories for prelabor rupture of membranes and the time of delivery.

Design: A retrospective cohort design within 2-year period (2015 to 2016) of a single-payer, government-owned health insurance database covering 75.8% individuals in a country

Setting: Nationwide healthcare providers (*n*=22,024) at primary, secondary, and tertiary levels

Participants: 12-to-55-year-old women that visit healthcare providers using the insurance from ~1% random sample of insurance holders stratified by healthcare provider and category of family: (1) never visit; (2) visit only primary care; and (3) visit all levels of care

Predictors: Medical histories of diagnosis and procedure (International Classification of Disease version 10) before the latest visit of outcome within the database period

Main Outcomes and Measures: Prelabor rupture of membranes prognostication (area under curve, with sensitivity, specificity, and likelihood ratio), the time of delivery estimation (root mean square error), and inference time (minutes), with 95% confidence interval

Results: We selected 219,272 women aged 33 ± 12 years. The best prognostication achieved area under curve 0.73 (0.72 to 0.75), sensitivity 0.494 (0.489 to 0.500), specificity 0.816 (0.814 to 0.818), and likelihood ratio being positive 2.68 (2.63 to 2.75) and negative 0.62 (0.61 to 0.63). This outperformed models from previous studies according to area under curve of an external validation set, including one using a biomarker (area under curve 0.641; sensitivity 0.419; sensitivity 0.863; positive likelihood ratio 3.06; negative likelihood ratio 0.67; *n*=1177). Meanwhile, the best estimation achieved ± 2.2 and 2.6 weeks respectively for predicted events and non-events. Our web application only took 5.14 minutes (5.11 to 5.18) per prediction.

Conclusions and Relevance: Prelabor rupture of membranes and the time of delivery were predicted by medical histories; but, an impact study is required before clinical application.

Keywords: prelabor rupture of membranes, preterm delivery, risk prediction, machine learning

Introduction

Preterm prelabor rupture of membranes (PROM) is widely used as an inclusion criterion for predictions of other conditions.[1-8] The disease precedes 40%~50% of all preterm deliveries and arises from multiple disease pathways.[9] Yet, the antecedent remains unclear, and prognostic predictions lack proper sample sizes and external validation.

Preterm delivery occurs in ~10% births in the United States, among which 2%~3% are contributed by preterm PROM, while the term PROM occurs in ~8% of pregnancies.[10] Premature babies require a neonatal intensive care unit (NICU),[11,12] which is scarce in several countries worldwide.[13-17] Meanwhile, use of NICUs accounts for a majority of healthcare costs in pediatrics and the single largest item in healthcare spending.[18] Predicting this disease, estimating the time of delivery, and tracing possible root causes would enable development of preventive strategies and improve efficiency of conducting a prospective cohort study of associated complications.

Prediction of PROM is mostly diagnostic.[19-22] For prognostications, a model of preterm PROM was recently developed using maternal factors during the first trimester.[23] Based on a training set (*n*=10,280), the area under receiver operating characteristic (ROC) curve (AUROC) was 0.667. Predictions of all-cause spontaneous preterm deliveries are also poor (AUROCs 0.54 to 0.70; <37 weeks' gestation; *n*=118/2540), and they are exposed to high risks of bias based on a systematic review.[24] However, there has been no development or external validation of a prognostic prediction model for PROM. In addition, because it is reasonably challenging, no studies have developed a model to estimate the time of delivery before the day.

For both classification and estimation tasks, machine learning algorithms have demonstrated promising performances for pregnancy outcomes[25] and other conditions.[26-28] Despite some hype, most of the greatest successes are diagnostic, especially deep-learning models that surpass human-level performance.[29-31] Machine learning, however, is yet unable to infer causality; thus, human learning is needed to estimate what will likely happen if conditions differ from those existing in the dataset from which a machine learns.[32] Nonetheless, solving PROM problems requires predictions and causal modeling to develop better preventive strategies at both the population and individual levels. To address

this issue, we applied both human and machine learning tools.[33] Using only medical histories, the model deployment could be accessible worldwide via a web application. This study aimed to develop, validate, and deploy a prognostic prediction model for PROM and an estimator of the time of delivery using a nationwide health insurance database.

Methods

We reported this study according to transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).[34] This study is a part of a DI-VNN project that applied our algorithm to various predicted outcomes. To develop the prediction models, we followed a protocol using the same hardware and software.[33] The checklists for all of the guidelines, including those in the protocol, and comparable models are available in eTables 1 to 5 in the Supplement. Ethical clearance was waived by the Taipei Medical University Joint Institutional Review Board (TMU-JIRB number: N202106025).

Study design

We applied a retrospective design to select subjects from a nationwide health insurance dataset provided by a government-owned health insurance company in Indonesia. The health insurance covered 200,259,147 (75.8%) individuals in that country,[35] including races of Asian and Austronesian. This dataset was the second version published on August 2019 (access approval no.: 5064/I.2/0421) covering ~1% (*n*=1,697,452) of insurance holders in 2015 and 2016 from nationwide, affiliated healthcare providers (*n*=22,024; primary, secondary, and tertiary care). Sampling for the data source was stratified by healthcare provider and category of family: (1) never visit; (2) visit only primary care; and (3) visit all levels of care. Details of these sampling procedures are described in the Supplement.

We included health insurance holders of 12~55-year-old females who had visited affiliated healthcare providers. We excluded visits after delivery. For a person who was pregnant twice within the period, we labeled the same person as a different subject for each pregnancy period. A complete list of codes for determining delivery or immediately after delivery care is available in eTable 7 of the Supplement.

We developed prediction models to classify if a visit was made by a subject for which the pregnancy period ended with PROM, and to estimate the time of delivery. Under-prognosis of PROM may cause pregnancy monitoring to be off-guard and was considered more serious than over-prognosis. A well-calibrated model with higher sensitivity should be given priority. Meanwhile, the error for estimating the time of delivery is acceptable around 2 to 4 weeks, since this is a common interval between antenatal visits closer to the time of delivery.

The outcome for the classification task was an event for a subject that had encountered the O42 code, which is the International Classification of Disease version 10 (ICD-10) code for PROM. Otherwise, a subject was assigned as a nonevent if the pregnancy ended within the dataset period using the same codes for pregnancy termination. If neither having pregnancy nor delivery, we assigned censoring labels. Based on the protocol,[33] we used these labels to preserve outcome distribution in the target population and resolve class imbalance by inversely weighting the uncensored outcomes considering both the censored and uncensored ones. Meanwhile, the outcome for the estimation task was the number of days from the latest visit encountering of the outcome code to a visit when the prediction model was used.

Candidate predictors consisted of medical histories defined by one or more codes of diagnosis and procedure. The multiple codes composed latent candidate predictors determined by model-based statistical tests to find potential candidate predictors based on systematic human learning.[36] Steps to determine candidate predictors were described in the protocol,[33] to avoid zero variance, perfect separation problem, outcome leakage, and redundancy, and to mimic real-world settings.[37,38] These resulted in 372 candidate predictors. Details of the candidate predictors and selection are described in eTable 8 of the Supplement.

Model development, validation, and deployment

Five models were developed. These are described in the protocol,[33] including hyperparameter tuning. Briefly, the first model was a statistical, i.e., ridge regression (RR), on 9 of 12 latent candidate predictors selected by the systematic human learning.[36] One of them was not associated with PROM (see Results), and we did not select low SES and maternal age for reducing use of private data and preventing social and economic discrimination. Opposite to systematic human learning, we applied unsupervised

machine learning to transform candidate predictors into principal components (PCs). These were used for supervised machine learning algorithms of an elastic net regression (PC-ENR), random forest (PC-RF), and gradient boosting machine (PC-GBM). The latter two algorithms outperformed other algorithms for pregnancy outcomes.[25] Based on the PC-ENR model, the PCs were reduced into 60 PCs (see the protocol[33]) to pursue 200 events per variable (EPV) for the PC-RF and PC-GBM, as recommended by the prediction model risk of bias assessment tools (PROBAST).[38] The fifth model was the DI-VNN. It was developed to achieve moderate predictive performance but interpretable results based on recent studies.[39,40] The DI-VNN allows deep exploration of how this algorithm works.[41] For this model, there were 144 candidate predictors after differential analyses with multiple testing corrections. For classification, we calibrated each model by a general additive model using locally weighted scatterplot smoothing (GAM-LOESS).

We split the dataset for internal and external validation. As recommended, we split out a dataset for external validation by geographical, temporal, and geotemporal splitting, approximately covering ~20% of visits. This reflected the situation in some real-world settings but not that in nationwide, which represented by ~20% random split of the remaining set; thus leaving ~64% of the original sample size for internal validation. We split out ~20% of the internal validation set to calibrate each model. The final predictive performance of internal validation came from this calibration subset by bootstrapping 30 times. Details on validation were described in the protocol.[33]

We deployed the best models as a web application. It needed a deidentified, two-column comma-separated value (CSV) file of admission dates and ICD-10 codes of the medical history of a subject. A user can set a threshold for the expected population-level sensitivity, specificity, positive predictive value, and negative predictive value. We showed an example in the Supplement. We also measured inference time for a prediction with 10 iterations.

Statistical analysis

All evaluation metrics are expressed as estimates with the 95% confidence interval (CI). Based 12 association diagrams (eFigures 1~12 in the Supplement) of PROM, we conducted inverse probability

weighting (IPW) to verify associations of the latent candidate factors and PROM (eTables 11~12 in the Supplement). Outcome regression was also conducted for comparison (eTables 13 in the Supplement).

Calibration (plot, intercept, and slope) and discrimination (i.e. AUROC) were computed for classification tasks. We also computed sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR-). For estimation, we computed a proportion of weeks in which each predicted time, in weeks, was included within an interval estimate of the true one. The interval had to be the maximum $\pm x$ weeks when predicting $> x$ weeks, e.g., for any women predicted to deliver in 6 weeks, this number should fall into the true time of delivery within ± 6 weeks. We determined the minimum and maximum predicted times of delivery with acceptable precision for each predicted outcome of PROM based on a visual assessment using internal validation. Root mean square error (RMSE) was computed within this acceptable range. We also evaluated the best time window using the best model for PROM prognostic predictions. The internal validation set was grouped by binning the days to the end of pregnancy every 4 weeks. An AUROC was computed for each bin. The best time window should be mostly greater than an AUROC of 0.5, which represents prediction by simple guessing.

Success criteria of the modeling were an AUROC greater than those of recent models (last 5 years) of PROM prognostic predictions using simple predictors (e.g., maternal factors), or greater or equal to those using high-resource predictors (e.g., biophysical or biochemical markers). To prevent a common cause of overfitting, we applied the PROBAST criterion, which is, the number of events in the training set should be ≥20 after being divided by the number of candidate predictors. We applied the preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 expanded checklist (eTable 4 in the Supplement) to find the comparable models.[42] Details on this procedure are described in the Supplement.

Results

We selected all visits (*n*=883,376) by 12~55-year-old women (*n*=219,272) (Figure 1 and Table 1). Of 12 latent candidate predictors, eleven had significant association with PROM, estimated by IPW (Table 2).

We depicted the final association diagram (eFigure 13 in the Supplement). This diagram and the association findings are further described in the Supplement.

Prognostic prediction of prelabor rupture of membranes (PROM)

After calibration (Figure 2a), the two well-calibrated models were the PC-ENR and DI-VNN; however, distributions of predicted probabilities from 0 to 1 were mostly covered by the latter model. For clinical application, this will help to widely adjust threshold, depending on the local data distribution. The optimal threshold for the DI-VNN was 0.14. Both models had visually differentiated distributions of predicted probabilities between events and nonevents. Weights, variable importance values, and intermediate outputs, which indicated the extent a predictor contributes to a prediction, are respectively shown for (1) the RR and PC-ENR (eTables 14 to 16 in the Supplement); (2) the PC-RF and PC-GBM (eTables 15, 17, and 18 in the Supplement); and (3) the DI-VNN (eTables 19 and 20 in the Supplement).

At 95% specificity, the PC-RF (0.513, 95% CI 0.509 to 0.517) was the most sensitive model (Figure 2b); unfortunately, this model was not well-calibrated. At the same specificity, the DI-VNN followed PC-RF with a sensitivity of 0.297 (95% CI 0.293 to 0.301; threshold at 0.29). With the optimal threshold at 0.14, the DI-VNN achieved a sensitivity of 0.494 (95% CI 0.489 to 0.5) and a specificity of 0.816 (95% CI 0.814 to 0.818). Potential utility of DI-VNN was also confirmed by LR+ (2.68, 95% CI 2.63 to 2.75) and LR- (0.62, 95% CI 0.61 to 0.63). By external validation (Figure 2c), the DI-VNN was the most robust.

For the random split that reflected common situations nationwide, the DI-VNN achieved an AUROC of 0.71 (95% CI 0.70 to 0.72). It was reasonably lower than that of a training set (0.73, 95% CI 0.72 to 0.75). For other external validations, the AUROCs of the DI-VNN were lower than that for the random split but higher than both the average AUROCs of all models and an AUROC of 0.5.

From PubMed, Scopus, and Web of Science, we identified 209 non-duplicated records. These were screened, retrieved, and assessed to find two prediction models[23,43] for preterm PROM (eFigure 14 and eTables 4 and 5 in the Supplement). Prognostic predictions of PROM by the DI-VNN and PC-RF achieved AUROCs that were higher than those of previous models (Figure 2b and 2c): (1) a logistic regression using maternal factors with an AUROC of 0.667 (sensitivity 0.25; specificity 0.90; LR+ 2.50;

and LR- 0.83) but without internal validation (144 preterm PROM; 10,136 not preterm PROM);[23] and (2) a prediction rule using serum alpha-fetoprotein with an AUROC of 0.641 (sensitivity 0.419; sensitivity 0.863; LR+ 3.06; and LR- 0.67) but without internal validation (31 preterm PROM; 1146 not preterm PROM).[43]

For an exploratory data analysis, most of the AUROC intervals were greater than 0.5 from 44 ± 2 weeks before the end of the pregnancy (eFigure 15a in the Supplement). Population-level data exploration is also extensively described in the Supplement. An interactive interface of the DI-VNN are provided in our web application (https://predme.app/promtime), allowing users to explore this model at both the population and individual levels.

Estimation of the time of delivery

Although term PROM may also happen, it is mostly related to preterm delivery. Since codes for preterm delivery or premature newborn might not be consistently assigned to all cases, we decided to estimate how many days from the current visit a mother would deliver. This estimation task is providing a benefit if the model could estimate the time of delivery within an interval estimate of a maximum $\pm x$ weeks when predicting $> x$ weeks. We found the highest 78.57% of the 42 weeks were precisely predicted by PC-RF according to those criteria (Figure 3a). The time estimated by the DI-VNN was unfortunately within only the 6^{th} week from the prediction dates of any sample. This was due to the differential analysis (see Methods), which filtered predictors based on categorical outcomes only. Nonetheless, we used the DI-VNN, as the best classification model, to stratify the estimated time of delivery, including that estimated by the PC-RF. We confirmed this model consistently outperformed the other models for the time estimation task using external validation sets. A comparison of estimation performances, including those by external validation sets, is further described in the Supplement.

With nonevents, the maximum estimated time of delivery by PC-RF was 36 weeks from a visit when the prediction was conducted, while the corresponding true time of delivery was 42 weeks (Figure 3b). This coincides with the maximum duration of the pregnancy. With events, the maximum predicted time was 6 to 10 weeks earlier.

Web application

A web application was provided using the best models. We hosted this application on a public repository for clinical prediction models (https://predme.app/promtime). This allows a user, e.g., a doctor, to make a decision after critically appraising an individual prediction by several approaches. We describe a case example in the Supplement, reflecting a real-world situation.

Discussion

The DI-VNN model in this study outperformed previous models[23,43]; it used a large training set and external validation sets (8778 visits and 3352 subjects for events only) and did not require biomarker testing. External validation by stratified random splitting was also applied to predict gestational diabetes using nationwide health insurance database.[44] The PC-RF model also estimated the time of delivery in weeks to predict preterm delivery, while previous models only predicted whether a preterm delivery would happen without estimating the date interval.[24] Our models also used a cohort paradigm to prevent temporal bias in the delivery prediction.[45] We did not only rely on the DI-VNN and PC-RF models but also datasets to estimate the model performances at the individual level based on subpopulations similar to that individual,[46] in a collaboration of machine learning and evidence-based medicine.[47] This framework also answers key challenges to evaluate model weaknesses,[48] verify if the predicted outcome is reasonable,[49] and make sensible clinical predictions by utilizing electronic medical records.[50]

As inferred from the PC-RF model, the time of delivery of PROM events was estimated earlier than that of nonevents. This disease may be considered an infection-related preterm delivery. Clustering analysis of placental gene signatures assigned this type of preterm delivery as a subclass of preeclampsia.[51] The DI-VNN model in this study also included medical histories of all subtypes of preeclampsia (eTable 20 in the Supplement). The DI-VNN model likely used preeclampsia-related codes as competing risks to predict PROM. Similarly, competing risk models were also previously developed for preeclampsia.[52] Exploratory data analysis in this study also demonstrated optimal predictive performances that cover a full pregnancy period and a month before the pregnancy. This is also similar to findings based on predictive modeling for preeclampsia.[25]

However, we noted several limitations in our study. Although the DI-VNN outperformed those from previous studies,[23,43] we could only apply it as a preliminary model for PROM because clinical acceptance requires an AUROC of ≥0.8.[53] More-sensitive models are still needed for use as second-line models. As recommended in a clinician checklist[53] for assessing the suitability of machine learning applications in healthcare (eTable 3 in the Supplement), external validation and determining an optimal threshold using local data are needed.

In conclusion, our DI-VNN model was found to be robust for PROM prognostic predictions. The PC-RF model was reasonably precise within a specific time window based on the predicted outcome by the DI-VNN. An impact study (i.e. clinical trial) are warranted to further assess these models.

Acknowledgments

The BPJS kesehatan in Indonesia gave permission to access the sample dataset in this study. This study was funded by the Ministry of Science and Technology (MOST) in Taiwan (grant number MOST109-2221-E-038-018 and MOST110-2628-E-038-001) and the Higher Education Sprout Project from the Ministry of Education (MOE) in Taiwan (grant number DP2-110-21121-01-A-13) to Emily Chia-Yu Su. These funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author contributions

HS, YWW, and ECYS developed the concept and design of this study (DBPR). Dataset access was requested by HS. This author extracted and processed the data, performed the training and validation of the machine learning algorithms, conducted the literature search, and wrote a draft of the manuscript. This author and YWW independently assessed the eligibility criteria of ambiguous, reviewed studies which were previously determined by HS. YWW and ECYS critically revised the draft manuscript. All authors approved the submitted manuscript and agreed to be personally accountable for their own contributions and to ensure the accuracy and integrity of any part of the work, including ones in which the author was not personally involved.

Competing interests

HS, YWW, and ECYS declare no competing interests.

Data availability

The data that support the findings of this study are available from the social security administrator for health or *badan penyelenggara jaminan sosial (BPJS) kesehatan* in Indonesia, but restrictions apply to the availability of these data, which were used under license for the current study (dataset request approval number: 5064/I.2/0421), and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the BPJS Kesehatan. To get this permission, one need to request an access from the BPJS Kesehatan for their sample dataset published in August 2019. Up to this

date, there are three sample datasets they published in February 2019, August 2019, and December 2020. For the first and second versions, a request is applied via https://e-ppid.bpjs-kesehatan.go.id/, while the third is applied via https://data.bpjs-kesehatan.go.id.

Code availability

The R Markdown, R Script, and others are available in https://github.com/herdiantrisufriyana/prom. To pre-process the raw data into the input dataset of this study, follow the codes of the R Markdown in https://github.com/herdiantrisufriyana/medhist/tree/main/preprocessing.

References

[1] Salman L, Aviram A, Holzman R, et al. Predictors for cesarean delivery in preterm premature rupture of membranes. *J Matern Fetal Neonatal Med* 2020;33:3761-66. doi: https://doi.org/10.1080/14767058.2019.1585422.

[2] Mikołajczyk M, Wirstlein P, Adamczyk M, Skrzypczak J, Wender-Ożegowska E. Value of cervicovaginal fluid cytokines in prediction of fetal inflammatory response syndrome in pregnancies complicated with preterm premature rupture of membranes (pprom). *J Perinat Med* 2020;48:249-55. doi: https://doi.org/10.1515/jpm-2019-0280.

[3] Cobo T, Munrós J, Ríos J, et al. Contribution of amniotic fluid along gestation to the prediction of perinatal mortality in women with early preterm premature rupture of membranes. *Fetal Diagn Ther* 2018;43:105-12. doi: https://doi.org/10.1159/000475926.

[4] Toukam ME, Luisin M, Chevreau J, Lanta-Delmas S, Gondry J, Tourneux P. A predictive neonatal mortality score for women with premature rupture of membranes after 22-27 weeks of gestation. *J Matern Fetal Neonatal Med* 2019;32:258-64. doi: https://doi.org/10.1080/14767058.2017.1378327.

[5] Esteves JS, de Sá RA, de Carvalho PR, Coca Velarde LG. Neonatal outcome in women with preterm premature rupture of membranes (pprom) between 18 and 26 weeks. *J Matern Fetal Neonatal Med* 2016;29:1108-12. doi: https://doi.org/10.3109/14767058.2015.1035643.

[6] Duncan JR, Dorsett KM, Vilchez G, Schenone MH, Mari G. Uterine artery pulsatility index for the prediction of obstetrical complications in preterm prelabor rupture of membranes. *J Matern Fetal Neonatal Med* 2019:1-4. doi: https://doi.org/10.1080/14767058.2019.1702961.

[7] Musilova I, Andrys C, Hornychova H, et al. Gastric fluid used to assess changes during the latency period in preterm prelabor rupture of membranes. *Pediatr Res* 2018;84:240-47. doi: https://doi.org/10.1038/s41390-018-0073-1.

[8] Sim WH, Ng H, Sheehan P. Maternal and neonatal outcomes following expectant management of preterm prelabor rupture of membranes before viability. *J Matern Fetal Neonatal Med* 2020;33:533-41. doi: https://doi.org/10.1080/14767058.2018.1495706.

[9] Menon R, Richardson LS. Preterm prelabor rupture of the membranes: A disease of the fetal membranes. *Semin Perinatol* 2017;41:409-19. doi: https://doi.org/10.1053/j.semperi.2017.07.012.

[10] Prelabor rupture of membranes: Acog practice bulletin, number 217. *Obstet Gynecol* 2020;135:e80-e97. doi: https://doi.org/10.1097/aog.0000000000003700.

[11] Braun D, Braun E, Chiu V, et al. Trends in neonatal intensive care unit utilization in a large integrated health care system. *JAMA Netw Open* 2020;3:e205239. doi: https://doi.org/10.1001/jamanetworkopen.2020.5239.

[12] Speer RR, Schaefer EW, Aholoukpe M, Leslie DL, Gandhi CK. Trends in costs of birth hospitalization and readmissions for late preterm infants. *Children (Basel)* 2021;8. doi: https://doi.org/10.3390/children8020127.

[13] Chellani H, Mittal P, Arya S. Mother-neonatal intensive care unit (m-nicu): A novel concept in newborn care. *Indian Pediatr* 2018;55:1035-36. doi, PMID: https://www.ncbi.nlm.nih.gov/pubmed/30745471.

[14] Umran RM, Al-Jammali A. Neonatal outcomes in a level ii regional neonatal intensive care unit. *Pediatr Int* 2017;59:557-63. doi: https://doi.org/10.1111/ped.13200.

[15] Shrestha D, Dhoubhadel BG, Parry CM, Prajapati B, Ariyoshi K, Mahaseth C. Predicting deaths in a resource-limited neonatal intensive care unit in nepal. *Trans R Soc Trop Med Hyg* 2017;111:287-93. doi: https://doi.org/10.1093/trstmh/trx053.

[16] Horbar JD, Edwards EM, Greenberg LT, et al. Racial segregation and inequality in the neonatal intensive care unit for very low-birth-weight and very preterm infants. *JAMA Pediatr* 2019;173:455-61. doi: https://doi.org/10.1001/jamapediatrics.2019.0241.

[17] Murthy S, Leligdowicz A, Adhikari NK. Intensive care unit capacity in low-income countries: A systematic review. *PLoS One* 2015;10:e0116949. doi: https://doi.org/10.1371/journal.pone.0116949.

[18] Ho T, Zupancic JAF, Pursley DM, Dukhovny D. Improving value in neonatal intensive care. *Clin Perinatol* 2017;44:617-25. doi: https://doi.org/10.1016/j.clp.2017.05.009.

[19] Julien M, N'Guema L, Bouzerara A, Toro B, Lecarpentier E, Guibourdenche J. Premature rupture of the membranes: Analytical evaluation of diagnostic tests. *Ann Biol Clin (Paris)* 2018;76:300-06. doi: https://doi.org/10.1684/abc.2018.1346.

[20] Eldaly A, Omran E, Youssef MA, et al. Use of beta subunit of human chorionic gonadotropin assay as a diagnostic tool for prelabor rupture of membranes. *J Matern Fetal Neonatal Med* 2019;32:1965-70. doi: https://doi.org/10.1080/14767058.2017.1422712.

[21] Ireland KE, Rodriguez EI, Acosta OM, Ramsey PS. Intra-amniotic dye alternatives for the diagnosis of preterm prelabor rupture of membranes. *Obstet Gynecol* 2017;129:1040-45. doi: https://doi.org/10.1097/aog.0000000000002056.

[22] Musilova I, Bestvina T, Hudeckova M, et al. Vaginal fluid interleukin-6 concentrations as a point-of-care test is of value in women with preterm prelabor rupture of membranes. *Am J Obstet Gynecol* 2016;215:619.e1-19.e12. doi: https://doi.org/10.1016/j.ajog.2016.07.001.

[23] El-Achi V, de Vries B, O'Brien C, Park F, Tooher J, Hyett J. First-trimester prediction of preterm prelabour rupture of membranes. *Fetal Diagn Ther* 2020;47:624-29. doi: https://doi.org/10.1159/000506541.

[24] Meertens LJE, van Montfort P, Scheepers HCJ, et al. Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: A systematic review and independent external validation. *Acta Obstet Gynecol Scand* 2018;97:907-20. doi: https://doi.org/10.1111/aogs.13358.

[25] Sufriyana H, Husnayain A, Chen YL, et al. Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis. *JMIR Med Inform* 2020;8:e16503. doi: https://doi.org/10.2196/16503.

[26] Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020;46:383-400. doi: https://doi.org/10.1007/s00134-019-05872-y.

[27] Lee Y, Ragguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord* 2018;241:519-32. doi: https://doi.org/10.1016/j.jad.2018.08.073.

[28] Gonem S, Janssens W, Das N, Topalovic M. Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax* 2020;75:695-701. doi: https://doi.org/10.1136/thoraxjnl-2020-214556.

[29] Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLoS Med* 2018;15:e1002699. doi: https://doi.org/10.1371/journal.pmed.1002699.

[30] Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019;25:65-69. doi: https://doi.org/10.1038/s41591-018-0268-3.

[31] Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Med* 2018;15:e1002686. doi: https://doi.org/10.1371/journal.pmed.1002686.

[32] Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020;2:e677-e80. doi: https://doi.org/10.1016/s2589-7500(20)30200-4.

[33] Sufriyana H, Wu YW, Su EC. Human and machine learning pipelines for responsible clinical prediction using high-dimensional data. *Protocol Exchange* 2021. doi: https://doi.org/10.21203/rs.3.pex-1655/v1.

[34] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. *Bjog* 2015;122:434-43. doi: https://doi.org/10.1111/1471-0528.13244.

[35] Ariawan I, Sartono B, Jaya C, et al. *Sample dataset of the bpjs kesehatan 2015-2016*. Jakarta: BPJS Kesehatan; 2019.

[36] Sufriyana H, Wu YW, Su EC. Systematic human learning by literature and data mining for feature selection in machine learning. *Protocol Exchange* 2021. doi: https://doi.org/10.21203/rs.3.pex-1634/v1.

[37] Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res* 2016;18:e323. doi: https://doi.org/10.2196/jmir.5870.

[38] Moons KGM, Wolff RF, Riley RD, et al. Probast: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med* 2019;170:W1-w33. doi: https://doi.org/10.7326/m18-1377.

[39] Sharma A, Vans E, Shigemizu D, Boroevich KA, Tsunoda T. Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci Rep* 2019;9:11399. doi: https://doi.org/10.1038/s41598-019-47765-6.

[40] Ma J, Yu MK, Fong S, et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* 2018;15:290-98. doi: https://doi.org/10.1038/nmeth.4627.

[41] Sufriyana H, Wu YW, Su EC. Deep-insight visible neural network (di-vnn) for improving interpretability of a non-image deep learning model by data-driven ontology. *Protocol Exchange* 2021. doi: https://doi.org/10.21203/rs.3.pex-1637/v1.

[42] Page MJ, McKenzie JE, Bossuyt PM, et al. The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Bmj* 2021;372:n71. doi: https://doi.org/10.1136/bmj.n71.

[43] Başbuğ D, Başbuğ A, Gülerman C. Is unexplained elevated maternal serum alpha-fetoprotein still important predictor for adverse pregnancy outcome? *Ginekol Pol* 2017;88:325-30. doi: https://doi.org/10.5603/GP.a2017.0061.

[44] Artzi NS, Shilo S, Hadar E, et al. Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* 2020;26:71-76. doi: https://doi.org/10.1038/s41591-019-0724-8.

[45] Yuan W, Beaulieu-Jones BK, Yu KH, et al. Temporal bias in case-control design: Preventing reliable predictions of the future. *Nat Commun* 2021;12:1107. doi: https://doi.org/10.1038/s41467-021-21390-2.

[46] Ng K, Kartoun U, Stavropoulos H, Zambrano JA, Tang PC. Personalized treatment options for chronic diseases using precision cohort analytics. *Sci Rep* 2021;11:1139. doi: https://doi.org/10.1038/s41598-021-80967-5.

[47] Scott I, Cook D, Coiera E. Evidence-based medicine and machine learning: A partnership with a common purpose. *BMJ Evid Based Med* 2020. doi: https://doi.org/10.1136/bmjebm-2020-111379.

[48] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;2:719-31. doi: https://doi.org/10.1038/s41551-018-0305-z.

[49] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347-58. doi: https://doi.org/10.1056/NEJMra1814259.

[50] Beam AL, Kohane IS. Big data and machine learning in health care. *Jama* 2018;319:1317-18. doi: https://doi.org/10.1001/jama.2017.18391.

[51] Leavey K, Benton SJ, Grynspan D, Kingdom JC, Bainbridge SA, Cox BJ. Unsupervised placental gene expression profiling identifies clinically relevant subclasses of human preeclampsia. *Hypertension* 2016;68:137-47. doi: https://doi.org/10.1161/hypertensionaha.116.07293.

[52] Wright D, Wright A, Nicolaides KH. The competing risk approach for prediction of preeclampsia. *Am J Obstet Gynecol* 2020;223:12-23.e7. doi: https://doi.org/10.1016/j.ajog.2019.11.1247.

[53] Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021;28. doi: https://doi.org/10.1136/bmjhci-2020-100251.

Table 1. Baseline characteristics of subjects for association tests and internal validation set.

Variable		Not PROM [a] (*n*=32,346)	PROM [a] (*n*=4,333)	*P* value
Pregnancy episode [b]	first pregnancy, [c] no. (%)	30,710 (94.94)	4,134 (95.41)	(reference)
	second pregnancy, [c] no. (%)	1,636 (5.06)	199 (4.59)	*P*=.19
Age	mean (SD), y	30 (6)	29 (6)	*P*<.001***
Insurance class	first, no. (%)	4,102 (12.68)	673 (15.53)	(reference)
	unspecified, no. (%)	94 (0.29)	17 (0.39)	*P*=.72
	second, no. (%)	10,802 (33.40)	1,698 (39.19)	*P*=.38
	third, no. (%)	17,348 (53.63)	1,945 (44.89)	*P*<.001***
Marital status	married, no. (%)	19,723 (61.0)	3,046 (70.30)	(reference)
	divorced/widowed, no. (%)	131 (0.4)	23 (0.53)	*P*=.57
	single, no. (%)	2,799 (8.6)	438 (10.11)	*P*=.81
	unspecified, no. (%)	9,693 (30.0)	826 (19.06)	*P*<.001***
Occupation segment	central-government-paid householder, no. (%)	10,378 (32.08)	913 (21.07)	(reference)
	employer householder, no. (%)	9,095 (28.12)	1,578 (36.42)	*P*<.001***
	employee householder, no. (%)	11,205 (34.64)	1,686 (38.91)	*P*<.001***
	unemployed householder, no. (%)	16 (0.05)	2 (0.05)	*P*=.64
	local-government-paid householder, no. (%)	1,652 (5.11)	154 (3.55)	*P*=.52
Multiple pregnancy	negative, no. (%)	32,143 (9.9e-01)	4,301 (9.9e-01)	(reference)
	positive, no. (%)	203 (6.3e-03)	32 (7.4e-03)	*P*<.001***
Chorioamnionitis	negative, no. (%)	32,333 (1.0e+00)	4,320 (1.0e+00)	(reference)
	positive, no. (%)	13 (4.0e-04)	13 (3.0e-03)	*P*<.001***
Intra-amniotic infection	negative, no. (%)	32,333 (1.0e+00)	4,328 (1.0e+00)	(reference)
	positive, no. (%)	13 (4.0e-04)	5 (1.2e-03)	*P*<.001***
Ante-partum hemorrhage	negative, no. (%)	32,207 (1.0e+00)	4,321 (1.0e+00)	(reference)
	positive, no. (%)	139 (4.3e-03)	12 (2.8e-03)	*P*<.001***
Genital tract infection	negative, no. (%)	32,328 (1.0e+00)	4,324 (1.0e+00)	(reference)
	positive, no. (%)	18 (5.6e-04)	9 (2.1e-03)	*P*<.001***
Periodontal disease	negative, no. (%)	32,209 (1.0e+00)	4,319 (1.0e+00)	(reference)
	positive, no. (%)	137 (4.2e-03)	14 (3.2e-03)	*P*<.001***
Polyhydramnios	negative, no. (%)	32,305 (1.0e+00)	4,326 (1.0e+00)	(reference)
	positive, no. (%)	41 (1.3e-03)	7 (1.6e-03)	*P*<.001***
Pneumonia	negative, no. (%)	32,332 (1.0e+00)	4,328 (1.0e+00)	(reference)
	positive, no. (%)	14 (4.3e-04)	5 (1.2e-03)	*P*<.001***
Asthma	negative, no. (%)	32,189 (1.0e+00)	4,311 (9.9e-01)	(reference)
	positive, no. (%)	157 (4.9e-03)	22 (5.1e-03)	*P*<.001***
Low socio-economic status	negative, no. (%)	14,982 (4.6e-01)	2,386 (5.5e-01)	(reference)
	positive, no. (%)	17,364 (5.4e-01)	1,947 (4.5e-01)	*P*<.001***
Maternal age	20 to 35 y, no. (%)	23,777 (7.4e-01)	3,401 (7.8e-01)	(reference)
	<20 or >35 y, no. (%)	8,569 (2.6e-01)	932 (2.2e-01)	*P*<.001***
Influenza	negative, no. (%)	31,451 (9.7e-01)	4,218 (9.7e-01)	(reference)
	positive, no. (%)	895 (2.8e-02)	115 (2.7e-02)	*P*<.001***

Percentages may not add to 100% due to individual rounding.
* *P*<0.05; ** *P*<0.01; *** *P*<0.001
[a] Subject per pregnancy episode (not including censored delivery)
[b] Not PROM vs. PROM (not including those who were not pregnant)
[c] The first and second pregnancies of a subject within the database period

Table 2. Association between each latent candidate predictor and PROM by inverse probability weighting.

Variable of interest	Unadjusted OR (95% CI; *P* value)	Adjusted OR (95% CI; P value)	Adjustment
Multiple pregnancy	1.038 (1.034 to 1.043; P<.001***)	1.062 (1.055 to 1.068; P<.001***)	Maternal age
Chorioamnionitis	1.394 (1.373 to 1.416; P<.001***)	1.351 (1.33 to 1.372; P<.001***)	Asthma + Influenza + Intra-amniotic infection
Intra-amniotic infection	1.119 (1.085 to 1.155; P<.001***)	1.118 (1.083 to 1.153; P<.001***)	Genital tract infection + Periodontal disease + Pneumonia + Multiple pregnancy
Ante-partum hemorrhage	0.924 (0.921 to 0.928; P<.001***)	0.929 (0.924 to 0.933; P<.001***)	Low socio-economic status + Maternal age
Genital tract infection	1.116 (1.101 to 1.132; P<.001***)	1.116 (1.101 to 1.132; P<.001***)	(no adjustment)
Periodontal disease	0.92 (0.917 to 0.922; P<.001***)	0.967 (0.96 to 0.973; P<.001***)	Asthma + Maternal age
Polyhydramnios	1.039 (1.029 to 1.049; P<.001***)	0.998 (0.989 to 1.006; P>.99)	Multiple pregnancy
Pneumonia	0.979 (0.97 to 0.987; P<.001***)	1.037 (1.025 to 1.049; P<.001***)	Asthma + Influenza
Asthma	0.958 (0.953 to 0.962; P<.001***)	0.971 (0.966 to 0.977; P<.001***)	Influenza
Low socio-economic status	0.979 (0.978 to 0.98; P<.001***)	0.979 (0.978 to 0.98; P<.001***)	(no adjustment)
Maternal age	0.969 (0.969 to 0.97; P<.001***)	0.969 (0.969 to 0.97; P<.001***)	(no adjustment)
Influenza	0.995 (0.993 to 0.997; P<.001***)	0.995 (0.993 to 0.997; P<.001***)	(no adjustment)

* *P*<0.05; ** *P*<0.01; *** *P*<0.001; CI, confidence interval

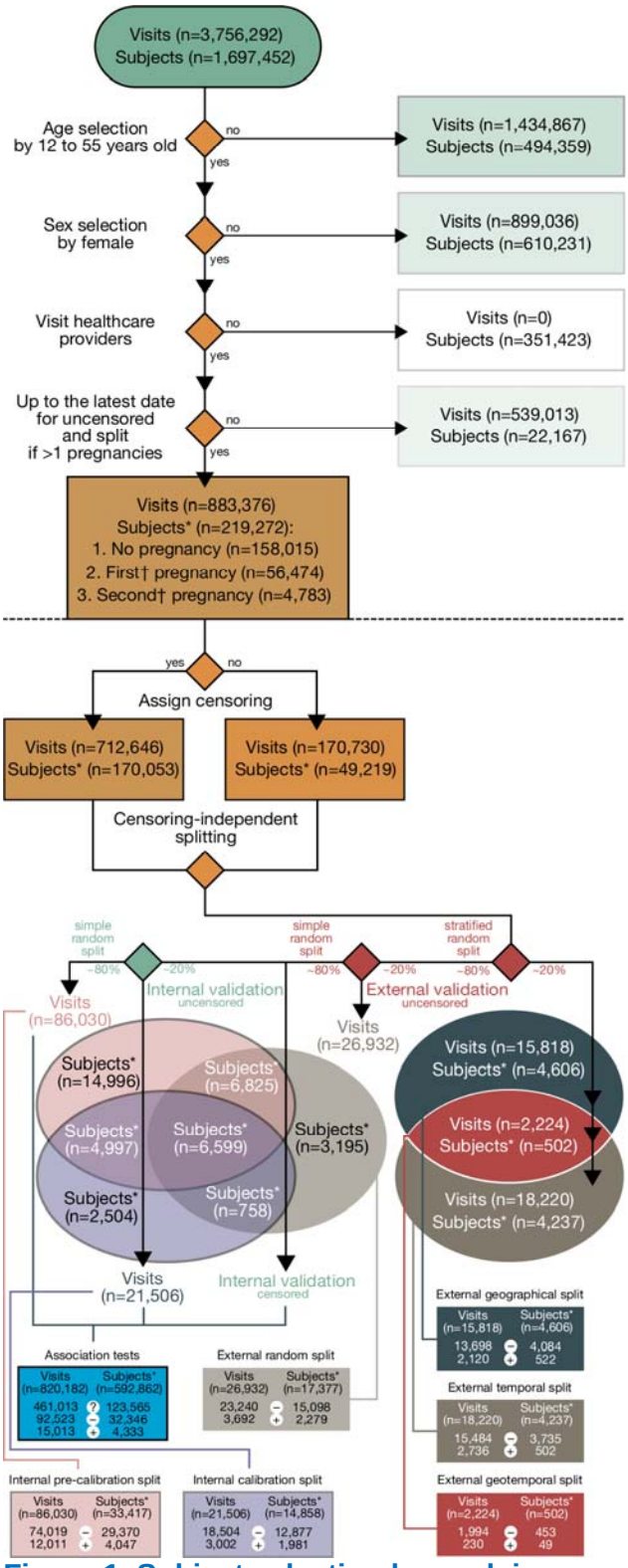

Figure 1. Subject selection by applying a retrospective design and data partitioning for internal and external validations. The set for association tests included censored outcomes. n, sample size; *, subject per pregnancy episode; **, the first and second pregnancies of a subject within the database period, not a parity; (?) censoring; (–) nonevents; and (+) events.

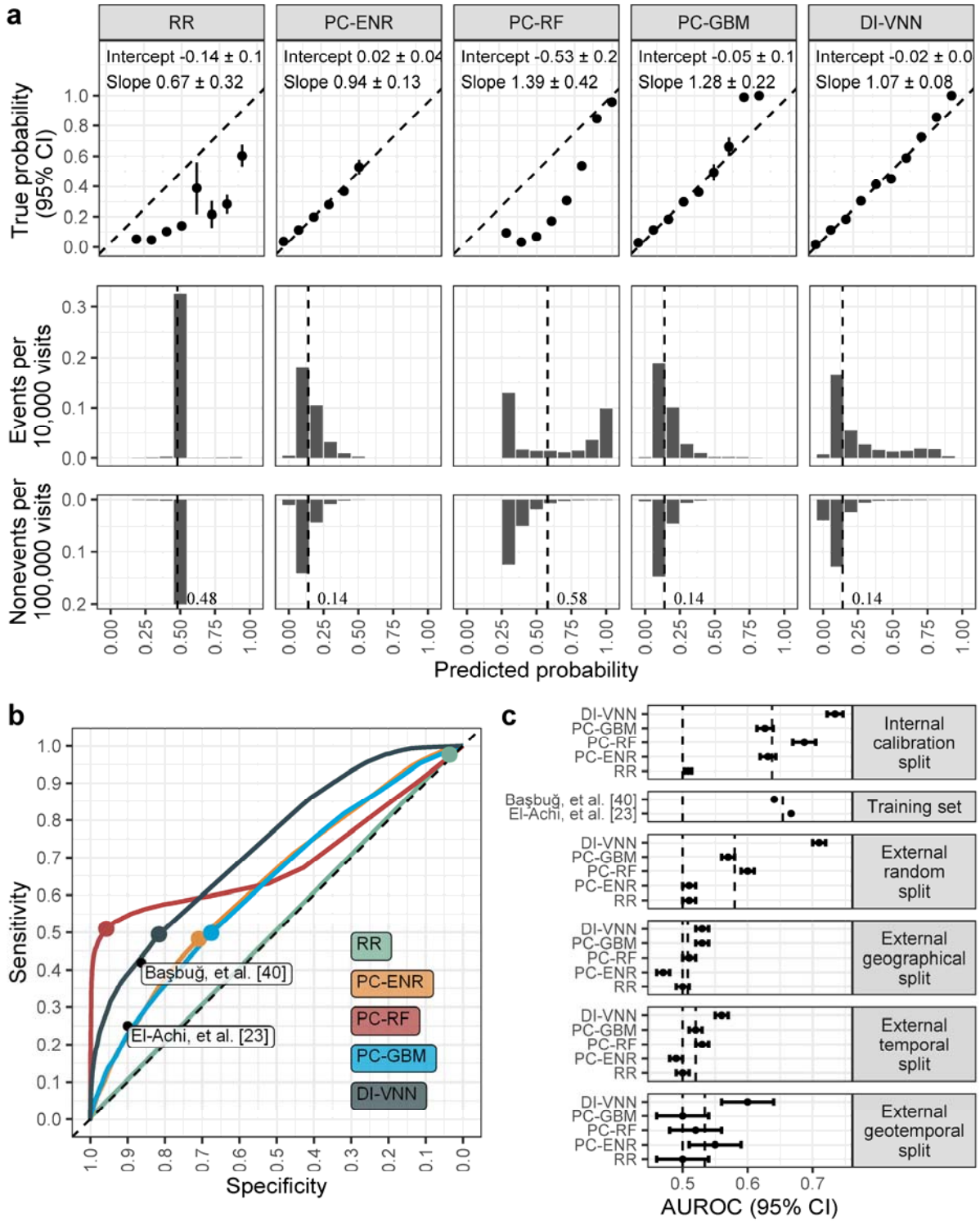

Figure 2. Model evaluation. (a) calibration; (b) receiver operating characteristic (ROC) curve; (c) area under ROC curve (AUROC). Thresholds (a, b) and average AUROCs per set (c). DI-VNN, deep-insight visible neural network; ENR, elastic net regression; GBM, gradient boosting machine; PC, principal component; RF, random forest; RR, ridge regression.

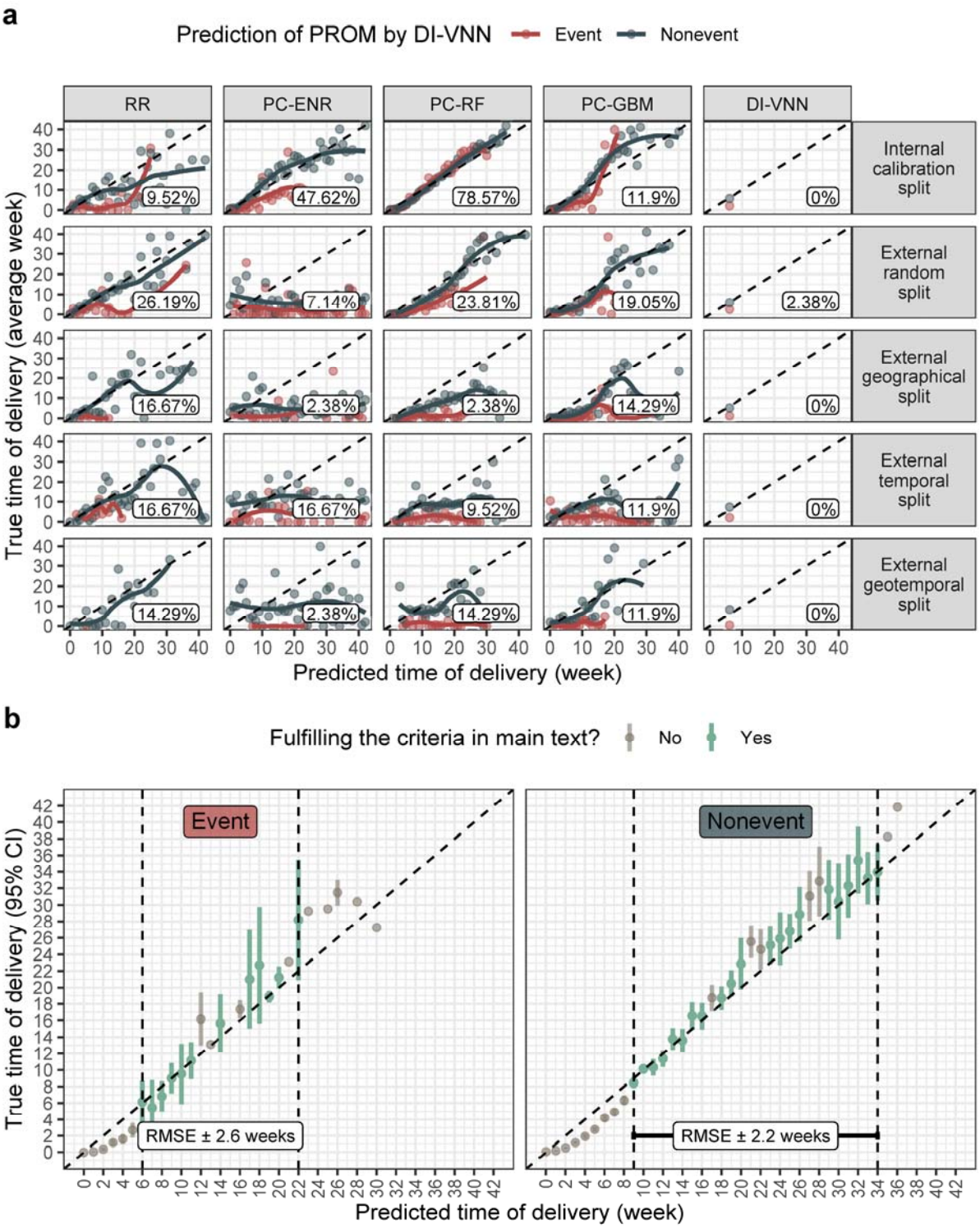

Figure 3. Estimation plots. (a) model comparison; (b) principal component-random forest (PC-RF) estimation window. Percent % criteria fulfilled (a) and precision (95% confidence interval) per week (b). DI-VNN, deep-insight visible neural network; ENR, elastic net regression; GBM, gradient boosting machine; RMSE, root mean squared error; RR, ridge regression.