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

Please kindly see List of Figures and List of eTables in Appendix.

# Introduction

In this Supplement, we describe details on this study following chronological order of our analysis pipeline on association test or predictive modelling for prelabor rupture of membranes (PROM). There are four sections corresponding to the same sections in the main text, which are respectively Introduction, Methods, Results, and Discussion. Along with this PDF document, we also provide R Markdown (.Rmd) containing the same texts with this document but including the programming codes for the data analysis in-between of these texts. An exception is subsection of “Comparison to previous studies”. The codes for core steps in the analysis pipeline are also provided exclusively in an R Script (.R). The codes beyond the core steps were used for analytic decision or creating tables or figures. These are shown to provide details on how data are processed to construct all tables and figures in both the main text and this Supplement, including those in Appendix of this Supplement (eTables 1 to 5, and 13) and those shown only in the R Markdown due to the complexity (eTables 6 to 12 and 14 to 21). A 5-minute video was provided in the protocol […] to briefly explain technical details on deep-insight visible neural network (DI-VNN) pipeline.

# Methods

## Research guidelines

To ensure that we conducted rigorous research, we carried out and reported this study by applying three sets of standard guidelines, specifically designed for a multivariable prediction model applying a machine learning algorithm that is suitable for healthcare (Luo W, et al, 2016; Scott I, et al, 2021; Moons KGM, et al, 2019). The checklists for all the guidelines are shown (eTables 1 to 3 in Appendix). For a fair comparison to previous models, we also followed methods in the preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 expanded checklist (Page MJ, et al, 2021). The checklist and the comparable models are also described (eTables 4 and 5 in Appendix). We also applied four approaches based on previous methods (Ritchie ME, et al, 2015; Maaten LVD, 2014; Ma J, et al 2018; Sharma A, et al, 2019; Stalpers LJA, et al, 2018; Goodman LA, 1961; Hernán MA, Robins JM, 2020; Lee J, Spratling R, 2019) to develop the pipeline with several modifications to provide a framework for improving interpretability of a deep learning model […]. All of the aforementioned procedures were comprehensively described in a human and machine learning pipelines we developed for responsible clinical prediction […].

## Programming environment

We set up a programming environment for this study. Bioconductor was utilized as described in the main text. There were 198 R packages which are 9 base packages, 53 other packages, and 136 dependencies (eTable 6 in R Markdown).

## Sampling procedures of the data source

The data source was a sample dataset of the whole health insurance database during 2015 and 2016 by cross-sectional design. Stratified random sampling was applied. The strata variable was constructed from 66,072 combinations of all the healthcare facilities (*n*=22,024) and category of family, which were: (1) a family of which members never visit the healthcare facilities; (2) a family of which members have visited only primary care; and (3) a family of which members have visited all levels of care. For each stratum, one to ten families were randomly included. This means only ten families were randomly included if more than that number, resulting 586,969 families with 1,697,452 subjects.

## The sampling procedures of the dataset in this study

We conducted non-essential data cleaning, e.g. revising the inconsistent name of states, estimating the healthcare identifiers, *et cetera*. These procedures were parts of our R package of medhist 0.1.0. No sampling was conducted.

After the non-essential data cleaning, we applied retrospective cohort design, as described in the main text. For pregnant women, we use several codes for determining delivery or immediately after delivery care. The 220 codes are described (eTable 7 in R Markdown).

## Data preprocessing

We conducted data pre-processing after defining the target population and sampling it retrospectively. Demographics were included as categorical variables for association tests. Then, we applied systematic human learning, as described in the main text, to determine what were latent candidate predictors that can be inferred from our dataset. We also computed a number of days for a code in the latest encounter before the time of prediction, including those by codes as a latent candidate predictor.

## Data partition

To ensure all inference or derivation using training set only, we need to conduct data partition before continuing the downstream analysis. We described data partition for model validation in the main text (see Methods).

## Association tests

We conducted association tests as described in the main text. This will help us to include only the confirmed associated predictors as latent candidate predictors before conducting pre-selection of those candidates to fulfill quality control of predictors in the main text. We included latent candidate predictors of which the data were available in training set. Details on this information and International Classification of Disease version 10 (ICD-10) codes or demographical variables for each candidate of latent candidate predictors are shown in the next section. A protocol of systematic human learning was followed for these association tests […].

## Quality control of candidate predictors

For determining candidate predictors, we utilized nationwide medical histories, excluding those for validation. But, for the downstream analysis, medical histories were provider-wise by estimation. This means our prediction models only used medical histories recorded by a healthcare provider, which was blinded to those recorded by others. This reflects most real-world situations in which a healthcare provider does not have access to medical records of other providers.

All candidate predictors, including non-demographical associated factors, have non-zero variances (eTable 8 in R Markdown). There were 460 candidate predictors fulfilling this criterion. We also showed in the same eTable that there are 426 candidate predictors without perfect separation

To prevent outcome leakage, we removed the maternal or baby diagnosis/procedure codes that demonstrated the delivery or after delivery care (typically up to 6 weeks following childbirth; eTable 9 in R Markdown). We only used the existing codes in the training set to determine outcome-leaker codes based on the previous codes for determining delivery or immediately after delivery care (eTable 7 in R Markdown). There were 54 codes that may leak the outcome.

To avoid redundancy, we computed pair-wise Pearson correlation coefficients (eTable 10 in R Markdown). None of the estimates showed a perfect correlation (*r*=1). High correlations (i.e., >~0.70) were reasonably identified between latent candidate predictors and the code components. We did not remove those pairs of predictors.

By systematic human learning and association tests using available data, we also determined latent candidate predictors (eTable 11 in R Markdown). There were 27 first- and 10 second-level factors that are considerably associated with PROM. Only data for 12 out of 27 factors were available in training set. Either the diagnosis/procedure codes, or demographical variables (not included as candidate predictors), for latent candidate predictors are also described (eTable 12 in R Markdown).

## Feature extraction as historical rates

To deal with a problem in which a healthcare provider does not have access to medical records of other providers, medical histories were quantified as Kaplan-Meier (KM) estimators for each candidate predictors, so called historical rates. We inferred these rates given the day number from a code encounter to current visit for each candidate predictor, derived only from the same nationwide medical histories, as previously described […]. This used irredundant candidate predictors with non-zero variances and no perfect separation in training set only. The candidate predictors were transformed into the historical rates in all data partitions.

## Feature representation as principal components by 10-fold cross validation

The historical rates of all candidate predictors were fitted to a principal component (PC) model. Only nationwide medical histories in training set were used for the model fitting. We applied 10-fold cross validation to estimate weights for all candidate predictors in each PC. A protocol for resampled dimensional reduction was followed for feature representation in this study […].

## Set up tuning-training-calibrating configuration and internal validation

Previous data partition had not held out instances for calibration yet. This took 80% of training set. We also gave different weights for event and non-event by including censored outcome, as described in the main text. For hyperparameter tuning, we applied 5-fold cross validation, instead of 10-fold as applied for PC modeling. Meanwhile, the final training and calibration for each model were conducted by bootstrapping for 30 times. The same resampling methods were applied for both classification and estimation tasks. Parallel computing by multiple central processing units (CPUs) were applied for training all models.

## Hyperparameter tuning, final training, and calibrating

We applied the tuning grids and the training configurations for all models, except DI-VNN which required several modifications. This is already described in a protocol we followed for these procedures, as previously described […]. More details for DI-VNN will be described in the next section.

## Deep-insight visible neural network (DI-VNN)

We applied different pre-processing pipelines for feature selection and representation in DI-VNN. Instead of PCs, we used the historical rates of the candidate predictors. A 5-minute video explaining DI-VNN pipeline is available in the protocol […]. Only pre-calibration training set was used for the downstream pipeline. We followed a protocol for this model […]. Briefly, we applied differential analysis for feature selection. Then, 1-bit stochastic gradient descent transformation was applied using the post-normalization, feature-wise average based on nationwide training set to determine if a value is lower, equal to, or higher than the average respectively as -1, 0, and 1. Using a feature-to-feature Pearson correlation matrix, we created a feature map *t*-stochastic neighbor embedding (*t*-SNE) with Barnes-Hut approximation. We also subset this map hierarchically by clique-extracted ontology (CliXO) algorithm based on the same correlation matrix. We constructed convolutional neural network (CNN) model using the hierarchy with input of feature map onto which transformed, candidate predictors were projected. We trained the CNN model to get the learning representation.

## Evaluating the best model for classification and estimation

Model evaluation is already clearly described in the main text. Practical costs of prediction errors were considered when evaluating the models. Under- prognosis may cause pregnancy monitoring to be off-guard. A pregnant woman with a preterm delivery may not reside in an area with a readily available NICU, particularly in low-resource settings. Over-prognosis may lead to unnecessary enrollment of patients into a cohort study in an early intervention for preventing PROM or complications, e.g., an antibiotic administration. This may also lead to unnecessary tests for more-specific prognostications.

In addition, unlike other time-varying outcomes, e.g., cancer, we did not predict a survival rate for the estimation task because the time interval for a pregnancy period is definitely known. It is also more intuitive for clinicians if the given information is the estimated days from the current visit to the day a pregnant woman will deliver a baby, as such normal delivery estimations are based on the last menstrual period and ultrasound examination.

We started from evaluating the calibration measures of all models for classification task. Then, we chose the best model for classification by AUROC using only internal validation (calibration split) bootstrapped for 30 times. Using the same subset, the best estimation model was also chosen.

## External validation

We also confirmed the robustness of the best models using external validation sets. By stratified random splitting, we provided three non-overlapping splits comprising ~20% for external validation: (1) geographical split; (2) temporal split; and (3) geotemporal split. We set these subsets challenging to predict by our models, in which the geotemporal split was the most difficult. This way we can estimate robustness of our model generalizability. In addition, uncertainty intervals were also computed by bootstrapping for 30 times. But, these do not reflect common situations nationwide; thus, the ~20% of the remaining was held out by simple random splitting for external validation, which is called external random split. For association tests and predictive modeling, including calibration, we only used the remaining ~64% of all the selected visits. A clear description on model validation is given in a protocol we followed for these procedures, as previously described […].

## Exploring the best model

Model exploration method is also described in the main text (see Model evaluation in Methods). In this Supplement (see R Markdown and R Script), we directly explored the best model for classification task. None of the remaining models were explored.

## Preparing web application

For web application, we prepared an example dataset, user interface, and processing script at the side of server computer. No line of codes for the web is included in the R Markdown or R Script. Description for this web application is already clearly described in the main text.

# Results

In the R Markdown, we show data for Figure 1 in the main text. Up to the latest date for uncensored outcome and after splitting if >1 pregnancies, the total visit was the sum of totals from all subsets, while the total subject was the sum of totals from all subsets with attention for the overlaps (d to j in footnote of data for Figure 1 in the R Markdown). For association tests, visits and subjects of the censored outcome were those after external random splitting (k and j in footnote of data for Figure 1 in the R Markdown). We also explicitly show PROM prevalence for each subset.

## Association diagram

For latent candidate predictors whose data were available, we created the association diagrams (eFigures 1 to 12 in Appendix). We excluded all common effects of PROM that we found during the systematic human learning. These are not needed for association tests. Instead, inclusion of these variables will cause collider-stratification bias. In the association diagrams, we apply different colors based on the types of nodes representing several factors: (1) type A is a first-level factor (with variable prefixed by A) that has a role as a confounder; (2) type I is a first-level factor (with variable also prefixed by A) that has a role as a candidate factor of interest; (3) type U is an unmeasured variable that can affect measure variable of type-A/I/Y variable; and (4) type Y is a target or dependent factor which is PROM. Node, of which variable denoted by asterisk, represents a type-A/I/Y variable that can be represented by several diagnosis or procedure codes. In each of the diagrams, we also show adjustment formulas which were used for association tests. Conceivably, all variables of type A/I/Y with asterisk may be included in each formula, but, some of these variables cannot be included because these were not available in our data, particularly in the training set (likely because of low prevalence). All measurement errors that may be affected by Type-U variables in this study were assumed as independent non-differential errors, because all data were measured from electronic medical records. As the main text, results of the association tests were also described (eTable 13 in Appendix), either by outcome regression or inverse probability weighting (IPW).

We would describe each association diagram. Fifty-six studies were found from PubMed (eTable 11 in R Markdown) […]. Since only some factors are possible to include in the association tests, only some of these studies were explained below. After verifying the assumption using our data, we constructed a final association diagram consisting all the confirmed latent candidate predictors.

### Multiple pregnancy

Association model of multiple pregnancy on PROM (ACOG, 2016a) included only maternal age as a confounder (Song J, et al 2019; Thilaganathan B, and Khalil A, 2014; Martin JA, and Osterman MJK, 2019). However, several confounders were not blocked yet: (1) Assisted reproduction (Lei LL, et al 2019; Thilaganathan B, and Khalil A, 2014); and (2) Race (Fiscella K, 1996; Martin JA, and Osterman MJK 2019). These are shown in the association diagram (eFigure 1 in Appendix). Assisted reproduction can be represented by diagnosis/procedure codes but unavailable in our training set. For race, it is conceivably not able being represented by those codes.

### Chorioamnionitis

There were two assumptions regarding relationship between chorioamnionitis and PROM. Chorioamnionitis may affect or be affected by PROM. We applied chorioamnionitis as the former one (eFigure 2 in Appendix), as previously demonstrated (Fukami T, et al 2017). Similar assumptions were also regarded between chorioamnionitis and intra-amniotic infection (IAI), but, we only treated chorioamnionitis as being affected by IAI (Tantengco OAG, et al, 2019). Except cigarette smoking (ACOG, 2016a; Kim CJ, et al, 2015), data for all confounders were available in the training set: (1) influenza (Littauer EQ, et al, 2017; Kim CJ, et al, 2015); (2) asthma (Baghlaf H, et al, 2019); and (3) IAI (ACOG, 2016a; Tantengco OAG, et al, 2019).

### Intra-amniotic infection (IAI)

Similar to chorioamnionitis, IAI may affect or be affected by PROM. Consistently, we treated IAI as the former one (ACOG, 2016a) (eFigure 3 in Appendix). We did not have data for these confounders, especially in training set: (1) cervical shortening (ACOG, 2016a; Kiefer DG, et al, 2009); and (2) race (Fiscella K, 1996; Menon R, et al, 2011). Therefore, we used these factors as the confounders: (1) genital tract infection (GTI) (Pandey D, et al, 2019; Yan JJ, et al, 2016; Romero R, et al, 2019; Tantengco OAG, et al, 2019); (2) periodontal disease (Figueiredo MGOP, et al, 2019; Stinson LF, et al, 2019); (3) pneumonia (Getahun D, et al, 2007; Stinson LF, et al, 2019); and (4) multiple pregnancy (ACOG, 2016a; Lee SM, et al, 2020).

### Ante-partum hemorrhage (APH)

Most data for confounders of ante-partum hemorrhage (APH) and PROM (ACOG, 2016a) were not available in the training set (eFigure 4 in Appendix). Only two confounders were adjusted: (1) low socioeconomic status (SES) (ACOG, 2016a; Bhandari S, et al, 2014); and (2) maternal age (Song J, et al, 2019; Fan D, et al, 2017). The other confounders were: (1) cigarette smoking (ACOG, 2016a; Bhandari S, et al, 2014); (2) illicit drug use (ACOG, 2016a; Bhandari S, et al, 2014); (3) race (Fiscella K, 1996; Shen JJ, et al, 2005); (4) assisted reproduction (Lei LL, et al, 2019; Qin J, et al, 2016); and (5) placenta on anterior wall (Torricelli M, et al, 2015; Fan D, et al, 2017).

### Genital tract infection (GTI)

Association between GTI and PROM (Pandey D, et al, 2019; Yan JJ, et al, 2016) was only confounded by tuberculosis (Fernández AA, et al, 2017; Sharma JB, et al, 2018). We did not have data for tuberculosis in the training set. Thus, GTI is the only variable in the association model (eFigure 5 in Appendix).

### Periodontal disease

We also found association between periodontal disease and PROM (Figueiredo MGOP, et al, 2019). An association model was constructed by adding these confounders: (1) asthma (Baghlaf H, et al, 2019; Moraschini V, et al, 2018); and (2) maternal age (Song J, et al, 2019; Genco RJ, and Borgnakke WS, 2013). Because of data availability, we could not include these confounders into the model: (1) stress (Wang W, et al, 2020; Genco RJ, and Borgnakke WS, 2013); (2) low education (Wang W, et al, 2020; Genco RJ, and Borgnakke WS, 2013); and cigarette smoking (ACOG, 2016a; Genco RJ, and Borgnakke WS, 2013) (eFigure 6 in Appendix).

### Polyhydramnios

Polyhydramnios was also an associated factor of PROM (Odibo IN, et al, 2016). The confounders were: (1) assisted reproduction (Thilaganathan B, and Khalil A, 2014; Lei LL, et al, 2019); and (2) multiple pregnancy (ACOG, 2016a; Moise KJ, 1997). Only multiple pregnancy data were available in our training set; thus, a PROM association model was constructed using polyhydramnios and multiple pregnancy (eFigure 7 in Appendix).

### Pneumonia

Association model of pneumonia on PROM (Getahun D, et al, 2007) was confounded by two factors. The first confounder was influenza (Littauer EQ, et al, 2017; Goodnight WH, and Soper DE, 2005), while the second one was asthma (Baghlaf H, et al, 2019; Goodnight WH, and Soper DE, 2005). Both were included in the causal model (eFigure 8 in Appendix).

### Asthma

As shown in the association model of pneumonia and PROM, asthma was also an associated factor of PROM (Baghlaf H, et al, 2019). Influenza was the only common confounder (Littauer EQ, et al, 2017; Murphy VE, et al, 2017). Therefore, we included this confounder in the association model of asthma on PROM (eFigure 9 in Appendix).

### Low socioeconomic status (SES)

Low SES was also indicated as an associated factor of PROM (ACOG, 2016a). We could not find the confounder. The association model only included low SES. We represented several demographical factors as low SES (eFigure 10 in Appendix).

### Maternal age

We could find maternal age as a confounder of PROM and multiple pregnancy/APH/ periodontal disease. Obviously, there is no confounder of PROM in an associated model with maternal age as the variable of interest (Song J, et al 2019). We included this variable exclusively in the association model (eFigure 11 in Appendix).

### Influenza

Similar to maternal age, obviously there is no confounder of PROM in an association model with influenza as the variable of interest (Littauer EQ, et al, 2017). This disease has a specific agent. Therefore, the PROM association model only included this disease (eFigure 12 in Appendix).

### Final association diagram

Chorioamnionitis (odds ratio [OR] 1.351, 95% confidence interval [CI] 1.33 to 1.372), intra-amniotic infection (OR 1.118, 95% CI 1.083 to 1.153), and genital tract infection (OR 1.116, 95% CI 1.101 to 1.132) were the top three highest effects estimated by IPW. Polyhydramnios was not verified as being associated with PROM by either IPW (OR 0.998, 95% CI 0.989 to 1.006) or the outcome regression (OR 1.238, 95% CI 0.851 to 1.801) using our data. Outcome regression showed the same ranks for chorioamnionitis and genital tract infection (GTI), while the effect estimate of intra-amniotic infection on PROM was not statistically significant by this method (OR 2.134, 95% CI 1 to 4.555). Three of 11 factors were assigned as associated factors by IPW but not by outcome regression. These included intra-amniotic infection and two variables: (1) pneumonia (OR 0.91, 95% CI 0.538 to 1.539); and (2) influenza (OR 0.957, 95% CI 0.863 to 1.061). Effects by outcome regression were mostly larger than those by IPW (eTable 13 in Appendix).

In the final association diagram (eFigure 13 in Appendix), infection- and immune-related conditions were seen: (1) influenza; (2) asthma; and (3) pneumonia. Maternal factors were also observed: (1) maternal age; (2) low socio-economic status; (3) multiple pregnancy; and (4) ante-partum hemorrhage. Both groups are shown colliding on periodontal disease in the final association diagram (eFigure 13 in Appendix), then collided with genital tract infection on intra-amniotic infection continuing to either chorioamnionitis or PROM.

## Prognostic prediction of premature rupture of membranes

Predictive performances for classification task were already clearly described in the main text. In this Supplement, parameter estimates in each model is described. However, these were not always straightforward because of the complexity of several models.

For ridge regression (RR), the estimates were weights or beta values, as commonly described in a regression model (eTable 14 in R Markdown). For classification task, top three highest weights were assigned for chorioamnionitis, IAI, and GTI. This is similar to the effect ranks in association tests by IPW.

Three models used principal components (PCs), including PC elastic net regression (PC-ENR). To transform predictors into a PC, each predictor has a weight to multiply with. These weights were also shown (eTable 16 in R Markdown). For the PCs themselves, the parameter estimates in PC-ENR were similar to those in RR. The weights in PC-ENR were also shown (eTable 16 in R Markdown).

For PC random forest (PC-RF) and PC gradient boosting machine (PC-GBM), the parameter estimates may be represented by variable importance. It is a proportion of learners (trees) that include a predictor. These numbers were also shown (eTables 17 and 18 in R Markdown).

For DI-VNN, we filtered predictors by differential analysis which consisted of multiple univariable linear regressions. The parameter estimates were expressed as log of fold changes. These were equivalent to log of odds ratios. This number and others, including false discovery rate (FDR), were also shown for the selected predictors by FDR <0.05 (eTable 19 in R Markdown).

Parameter estimates of the DI-VNN were extremely enormous. To get similar sense with those of the other models, we used an intermediate output at a layer after being fed to Inception v4 for each ontology. This showed a learning representation by DI-VNN on a predictor. How these were computed had been described in a protocol we followed for these procedures […]. The intermediate outputs were shown (eTable 20 in R Markdown). We would point out several meanings of these outputs after the next section. In addition, connections between ontologies under the root are also shown (eTable 21 in R Markdown).

For comparison of our models with those of previous studies, we applied methods in the preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 expanded checklist (eTable 4 in Appendix). From three literature databases and several steps, we found two prediction models as described in the main text (eTable 5 in Appendix). The steps are shown (eFigure 14 in Appendix).

## Estimation of the time of delivery

Estimation performances were already clearly described in the main text. Parameter estimates of the estimation models were also shown in the same eTables with those of the classification models (eTable 14 in R Markdown). For  
estimation task, which implicitly related to preterm delivery, the lowest (negative) weights in the RR were also assigned to chorioamnionitis, IAI, and  
GTI, but, the second lowest rank was shifted up by multiple pregnancy. Lower- rank weight means earlier time of delivery, which is, likely having more chance to be a preterm delivery. It turns out our decision to choose IAI and chorioamnionitis as the associated factors, instead of those being affected of  
PROM, is consistent with these findings.

Although the RR mostly had achieved highest proportion of criteria fulfilled among the models by external validation sets, we noticed this only applies for  
the predicted non-events based on visual assessment. The associated factors may be well-generalized to estimate the time of delivery under non-events predicted (modeled) by DI-VNN. Nevertheless, we selected PC-RF as the best model for estimation task of the time of delivery.

Challenge on estimation task is reasonable considering different distributions among internal and external validation sets. In internal validation set, of which we only utilized the calibration set for model evaluation, predicted events by DI-VNN happened 15 weeks (95% CI 11 to 18; *n*=760) on average from the time of prediction. This was similar to those in external random (18, 95% CI 14 to 21; *n*=973) and temporal splits (15, 95% CI 12 to 19; *n*=687), but later than those in external geographical (9, 95% CI 6 to 13; *n*=500) and geotemporal splits (3, 95% CI 1 to 5; *n*=157). Meanwhile, the pregnancies predicted as nonevents by DI-VNN ended at 40 weeks (95% CI 39 to 41; *n*=20,746) on average from time of prediction. This was earlier than those in external random (42, 95% CI 40 to 43; *n*=25,959), temporal (52, 95% CI 50 to 53; *n*=17,533), and geotemporal splits (60, 95% CI 54 to 65; *n*=2,067), but later than that in external geographical split (37, 95% CI 35 to 38; *n*=15,318).

## Exploring deep-insight visible neural network

Population-level data exploration is described in this Supplement. For interactive figure and table of DI-VNN, we provide these in our web application (<https://predme.app/promtime>). Technical details for exploring DI-VNN were already described in a protocol we followed for these procedures […].

The network architecture of the DI-VNN is data-driven (eFigure 15b in Appendix). Using the clique-extracted ontology (CliXO) (Kramer M, et al, 2014) algorithm, we constructed an architecture for the convolutional neural network (CNN) using only a training set (see Methods). We can consider this algorithm as an agglomerative hierarchical clustering algorithm, but there could be more than one parent-child connection. A child ontology array has a similar value distribution with the parent one is because a neural network applies a backpropagation algorithm that updates the model parameters consecutively from the surface to the deeper layer following the path, which is the ontology network of DI-VNN. But, unlike other neural network models, DI-VNN isolates the backpropagation effect; therefore, we can trace the array values to interpret the possible meaning.

We explored each node of the DI-VNN at the population level. The diagnosis/procedure codes constructing the feature members were ICD-10 codes. We began from the most visually distinguished array (eFigure 15b in Appendix) which was ONT:171. These were N760 (acute vaginitis) and B379 (unspecified candidiasis). A positive output does not necessarily refer to an event. Nevertheless, positive and negative (color-coded) outputs tend to contribute to opposite outcomes, which are interpreted based on external contextual knowledge. Another distinguished array, ONT:144, was connected to the same node as that of the previous array. The feature member (9059, other microscopic examination of blood) tended to contribute to the same outcome as that of “unspecified candidiasis”. Unlike “acute vaginitis”, which is a local infection, both B379 and 9059 may be related to systemic infections. To this point, we gained insights into describing coincidences between systemic vs.  local infection and PROM. Another array, ONT:154, was also visually distinguished. Although the value of the feature member causal\_A03 (chorioamnionitis) was zero, it was next to higher values that supported the same outcome as that of “acute vaginitis”. We traced the node on the upper layer to ONT:171. This array had a similar value distribution on the same channel (*z*=2) to that of ONT:154.

The feature position within any ontology array was determined using *t*-SNE with the Barnes-Hut approximation (see Methods) (Maaten LVD, 2014). It mapped features on high dimensional to lower dimensional space, as multiple localities. This algorithm spreads small clusters instead of making a large bubble of clusters; thus, we expected our CNN algorithm can be better extracting predictive features from these localities. If a feature nearer to one than another, this means there is a closer relationship between both features. The localities clustered by *t*-SNE are grouped together at the root node on the most superficial layer (eFigure 15b in Appendix). Deeper layers have different subsets of features separated by the ontology grouping. By understanding how this algorithm works, we assumed causal\_A03 (ONT:154), i.e., “chorioamnionitis”, was closer to “acute vaginitis” than was “unspecified candidiasis” since both were in the same channel, the positions of the first and second dimensions were adjacent, and *t*-SNE preserved the neighborhood identity. Acute vaginitis is semantically a genital tract infection (GTI). By external contextual knowledge inferred from systematic human learning and confirmed by IPW using our data, GTI and chorioamnionitis were determined to be associated factors of PROM.

A node on more superficial layer, which is ONT:155, consisted a feature that prefers the same outcome that N760 (acute vaginitis) and causal\_A03 (chorioamnionitis) prefer, which is 598 (urethral catheterization). This feature is semantically related to acute vaginitis because the anatomical sites are adjacent. But, since ONT:155 is on a more superficial layer, this node will connect to the same node with many features from other ontology terms. This means more factors may need to interact with urethral catheterization (598) to be predictive for PROM. In addition, within the same ontology term, there is also 8602 (injection or tattooing of skin lesion or defect). Apparently, the CliXO algorithm have clustered these features together semantically, which are similarly invasive procedures. In addition, local infection from acute vaginitis may be related to chorioamnionitis, such that prefers the same outcome, as opposed to the possible systemic infection by the unspecified candidiasis.

Exploring this DI-VNN should be done with caution, since we developed the model using medical histories, which are diagnosis or procedure codes provided by medical doctors. We may or may not be modeling a human pathophysiology, but, we are definitely modeling the doctors’ behaviors of coding a diagnosis or procedure (Beaulieu-Jones BK, et al, 2021). By providing an interface to the internal properties of this model, a human user can assess each prediction case-by-case. From ONT:167 and ONT:149 on the deeper layer, we can find unusual features in the context of PROM, which are H527 (unspecified disorder of refraction) preferring nonevents while 734 (flat foot), H521 (myopia), and H522 (astigmatism) preferring events. These codes might be responses to the subject symptoms of edema in the feet and blurry vision. Both symptoms in a pregnant woman may be typically associated with severe preeclampsia. But, a doctor may avoid this association if the context does not support the symptoms, e.g. symptoms by a non-pregnant subject. This may lead a doctor to assign these codes responding to those symptoms. For each prediction, a human user may need to explore the model to avoid misclassification by ignoring the prediction if it counters the clinical reasoning. More pragmatically at individual level, the predictive value may not be sufficient or the estimation may not be quite precise based on the corresponding subpopulation data with respectively the same predicted outcome or estimated time of delivery. In addition, albeit all of the population-level patterns from this exploration, every subject may reveal a different pattern using the same DI-VNN model, as described in the next section.

## Web application

Briefly, we uploaded a record of 20 visits by a 19-year-old female subject from December 2, 2015 to July 30, 2016, consisting of 28 code entries. After determining the prediction date, which was set to July 30, 2016, we ran the application in 5.14 minutes (95% CI 5.11 to 5.18 minutes when we repeated it 10 times). We downloaded the report after the application was completed (eFigure 16 in Appendix). The predicted outcome was PROM, and the estimated time of delivery was 11 weeks after the time of prediction. The predicted probability was 0.867.

Furthermore, after the application was done (eFigure 16 in Appendix), we tried to adjust the threshold to the maximum value, such that the data for population-level performances are still available and the predictive value is also maximized depending on the prediction result. The population-level data was the same with internal validation (calibration split; *n*=21,506) which was used to plot the ROC curve (Figure 2b in the main text). The threshold was 0.67 with positive predictive value of 0.809 (95% CI 0.798 to 0.82). The sensitivity was reasonably low (0.107, 95% CI 0.104 to 0.11) if using this threshold. But, from a standpoint of prediction at individual level, a precise estimation is important to determine whether a decision corresponding to this prediction can be made with a good confidence. By default, the threshold is set at an optimum value of 0.14 (sensitivity 0.494, 95% CI 0.489 to 0.5). Based on the predicted probability case-by-case, a user can decide the threshold to adjust at.

From the reported timeline (eFigure 6 in Appendix), this subject is shown predicted to deliver on October 18^th 2016, approximately. If the predicted outcome is not PROM, the shaded area would be red; otherwise, turquoise color is applied to the area as shown. It depicted population-level estimation of true time of delivery for subjects that were also estimated to deliver within 11 weeks and predicted as PROM by the same threshold. The population-level data was the same with internal validation (*n*=107,536) which was used to plot the PC-RF estimation window (Figure 3b in the main text). By population-level estimation, the time of delivery might be at the beginning up to the end of October. Using threshold at 0.67, this population-level estimation was shifted earlier for a week compared to that at 0.14. In addition, for illustration purpose, just like a real-world setting, say we know the gestational age was 22-23 weeks’ gestation based on last menstrual period and ultrasound examination. If this estimation model is precise for this case, the subject would deliver at 33-34 weeks’ gestation, which is a preterm PROM.

We also saw the medical history of this subject from the reported timeline (eFigure 16 in Appendix). Up to the date of prediction, the prediction model used these features: (1) A09 (diarrhea and gastroenteritis of presumed infectious origin); (2) J069 (unspecified acute upper respiratory infection); (3) K30 (dyspepsia); and (4) 8878 (diagnostic ultrasound of gravid uterus). On the timeline, these were ordered from the most positive (top) to the most negative (bottom) based on each output in the ontology array.

In the prediction model, any of these features were classified in the ontology arrays of ONT:158, ONT:169, ONT:176, and root, as depicted by the ontology network (eFigure 16 in Appendix). We also identified the deepest ontology that includes all features in the timeline, which is ONT:169, but the predicted outcome based on this ontology is not PROM using the same threshold with that of the root. A user also can see a predictive performance of any ontologies, just like AUROC of the root (0.714, 95% CI 0.712 to 0.716). It is computed for the pre-calibrated DI-VNN only, since the calibrated one only used the predicted probability that was the output convoluted from the root ontology array.

Negative values at population level tend to event, as described in the previous section. From the ontology array (eFigure 16 in Appendix), J069 (unspecified acute upper respiratory infection) tends to event in that array, as shown as mostly negative outputs in the timeline. This feature was also surrounded by more negative outputs. A09 (diarrhea and gastroenteritis of presumed infectious origin) also had a negative value, but, this feature and K30 (dyspepsia) were surrounded by more positive outputs in the ontology array. If we apply the same illustrative gestational age, these infectious diseases (J069 and A09) were diagnosed respectively ten and four weeks the start of pregnancy, as shown on the timeline. Root is the only ontology that predicted PROM and included all features in this subject. This is implied all of these features should be taken together for the prediction. In addition, one may question why J069 (unspecified acute upper respiratory infection) tends to event while influenza have the opposite effect (OR 0.995,95% CI 0.993 to 0.997; eTable 13 in Appendix). We found that influenza, which is causal\_A28, did not include J069 (eTable 12 in R Markdown). This implied specific acute upper respiratory infection, such influenza, may not have the same effect with that by the unspecified one on PROM.

Beyond the root, the array of ontology ONT:169 is also shown (eFigure 17 in Appendix). A09 and J069 had negative values. As described in the main text, these features tend to an event. Respectively, the surrounding positive and negative values were subtler in this ontology array. Since this filtered array is fed forward to convolutional layers to be reduced each time passing a layer (see video in the protocol [a protocol citation]), the value of A09 and J069 were averaged along with the adjacent values toward zero, opposite to event. In turn, this coincides with lower AUROC in ONT:169 compared to that in root.

Eventually, a user may want to know if the PROM prediction and estimated time of delivery are similar to true values. One can save this model online and return later to the web application to enter the true outcome and time of delivery. In this way, a user can collect data for external validation purposes, specifically describing the model performances based on local data distributions. In our case, the true outcome was also PROM and the time of delivery was 12 weeks after the time of prediction, a week later than the predicted time of delivery.

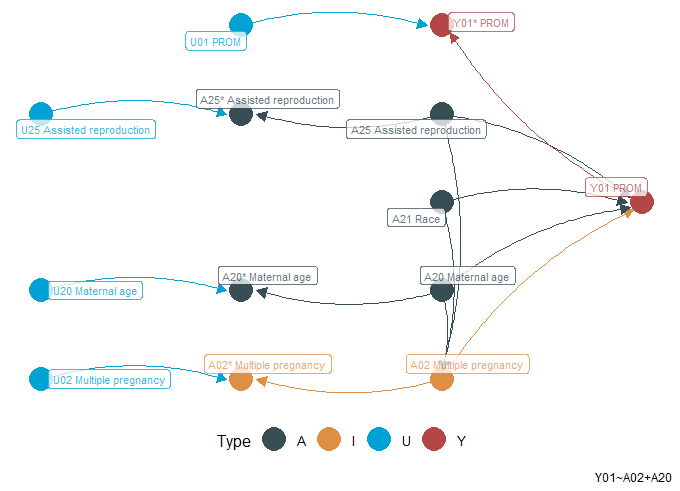
# Discussion

From association tests and feature extraction to model selection and exploration, we only used an internal validation set. But, the model is evaluated using four external validation sets with a large sample size. This has found the DI-VNN prediction was robust and the PC-RF estimation was precise within a reasonable time window. All of these models used only a medical history of a patient, which is easily extracted from the electronic medical record systems of most healthcare providers worldwide. Neither a biomarker testing nor even a laboratory test is needed. Eventually, the best models are ready to use for any healthcare providers using an open-access web application without changing their electronic health record systems and revealing private data.

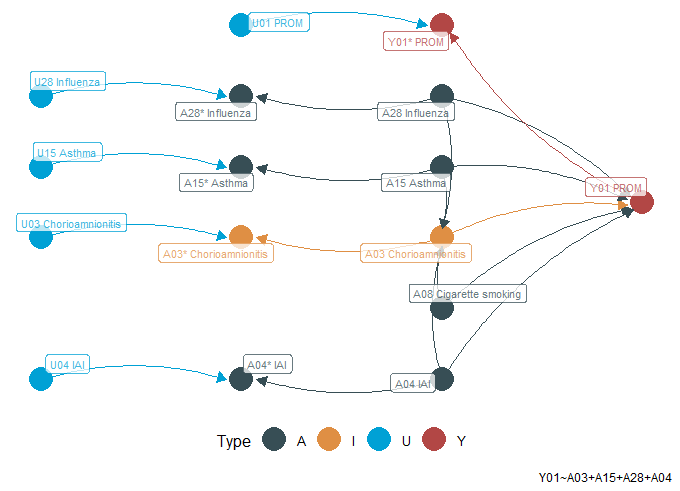
To gain insights into PROM antecedents, we showed how this was achieved by exploring the model at both the population (eFigure 15b in Appendix) and individual levels (eFigure 16 in Appendix). If we assume that PROM is a subclass of preeclampsia, this may explain why population-level exploration implied an insight of systemic vs. local infection by competing risks. A hematogenous infection may be associated with reproductive-tract microbial dysbiosis and can affect several pregnancy outcomes, including PROM, preeclampsia, and fetal growth restriction (Amir M, et al, 2020). Both hematogenous and ascending infections from reproductive tract are found in PROM (Romero R, et al, 2019; Solt I, 2015). Hematogenous infections included those from digestive and respiratory organs. Similar to the population-level exploration (eFigure 15 in Appendix), we also found a possibility of hematogenous infection at the individual level (eFigure 16 in Appendix), and these were from infectious gastroenteritis and unspecified acute upper respiratory infection. Both population- and individual-level explorations also implied a period surrounding the beginning of pregnancy as the onset of the optimal time to predict PROM. Several features in the DI-VNN were counterintuitive, e.g., H527 (unspecified disorder of refraction), 734 (flat foot), H521 (myopia), and H522 (astigmatism). Yet, these describe blurry vision and swelling in the feet, which are also symptoms of preeclampsia. Similar eye-related codes, i.e., myopia and astigmatism, were also found to be important in predicting preeclampsia in an RF model (Sufriyana H, et al, 2020).

# References

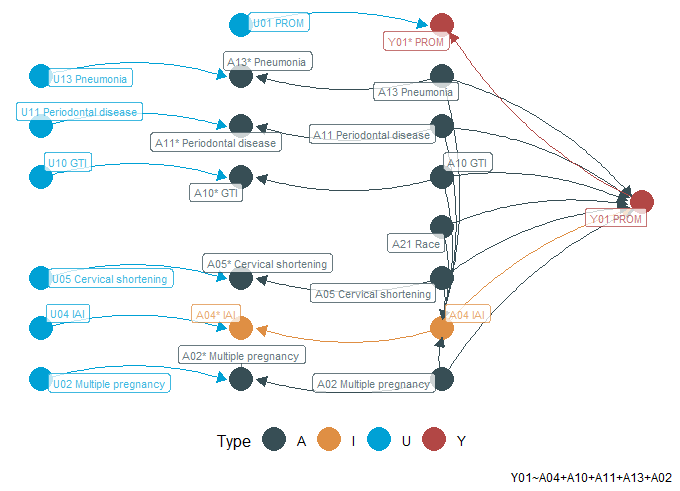
# Appendix



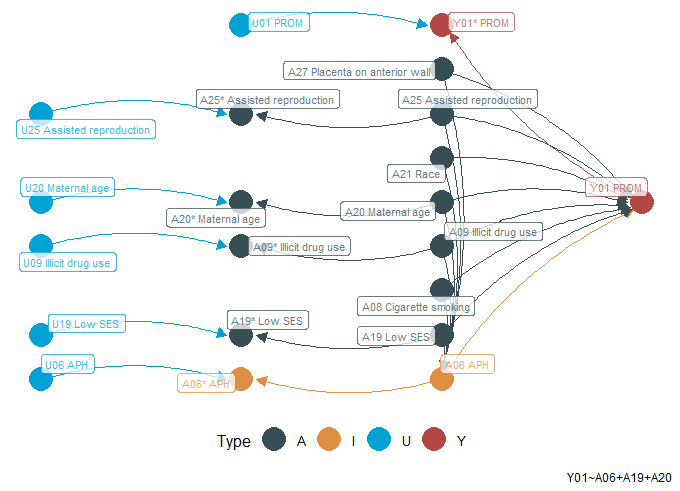
Multiple pregnancy



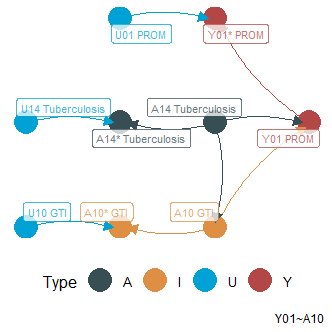
Chorioamnionitis



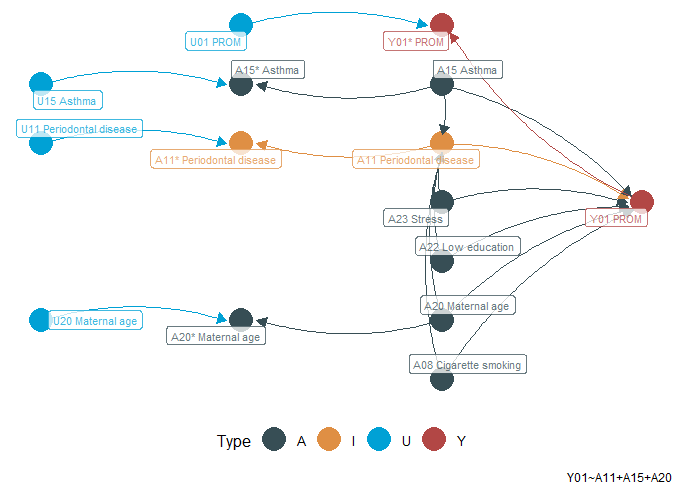
Intra-amniotic infection (IAI)



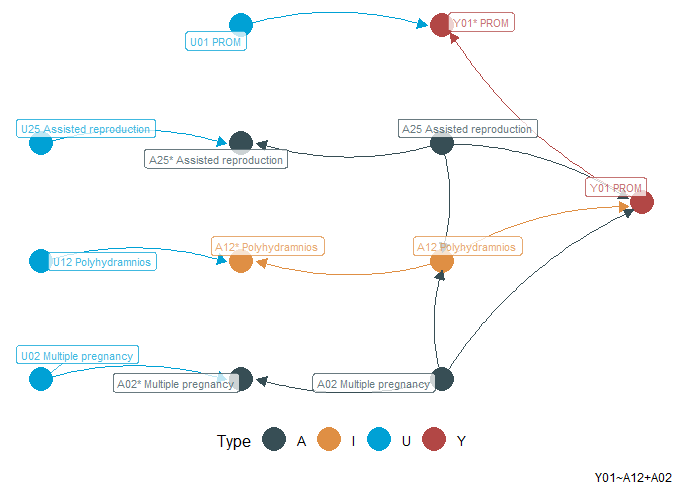
Ante-partum hemorrhage (APH)



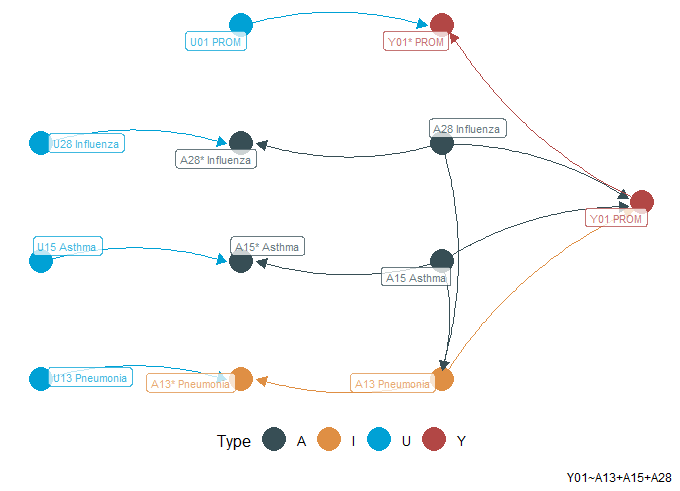
Genital tract infection (GTI)



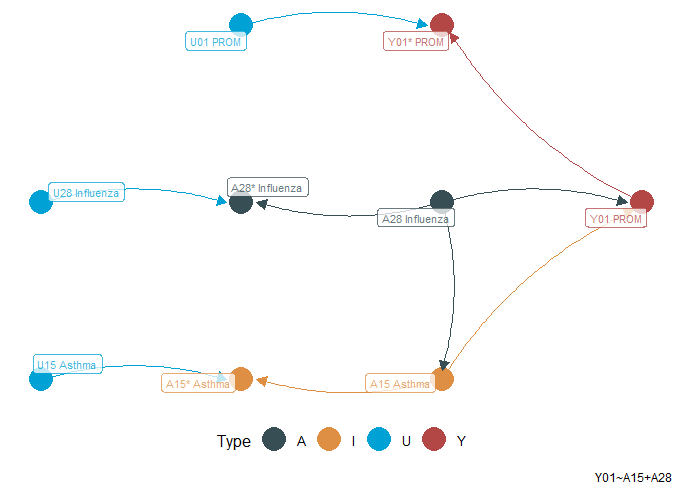
Periodontal disease



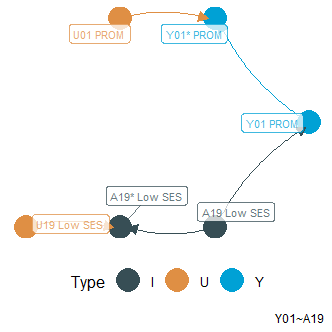
Polyhydramnios



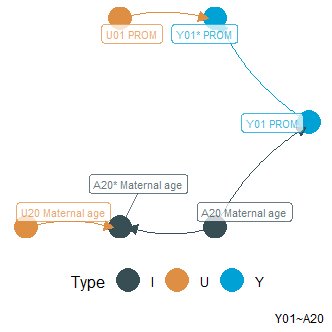
Pneumonia



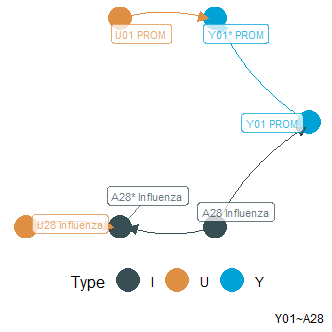
Asthma



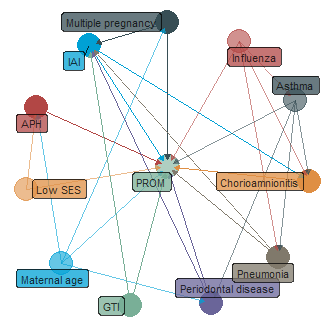
Low socio-economic status (SES)



Maternal age



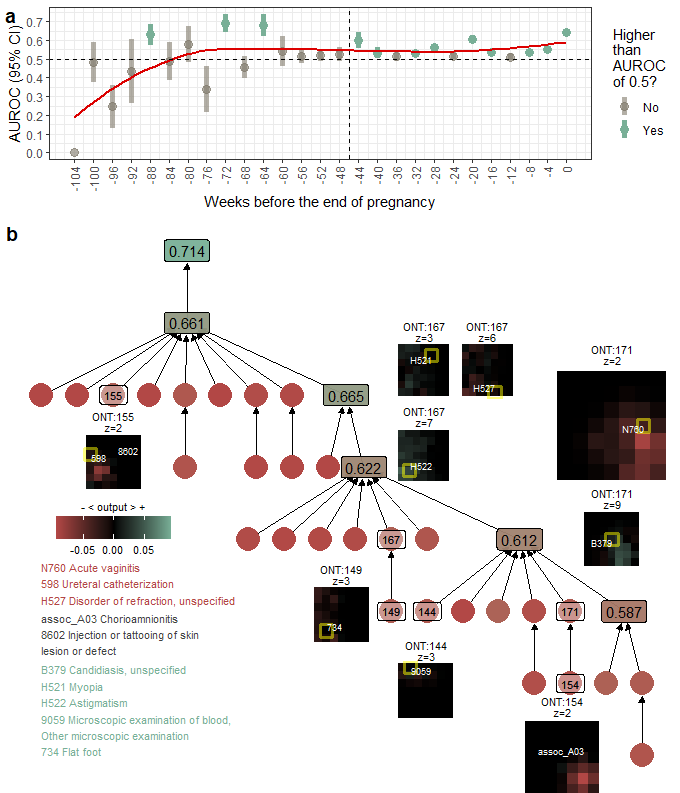
Influenza



Final association diagram

Caption:

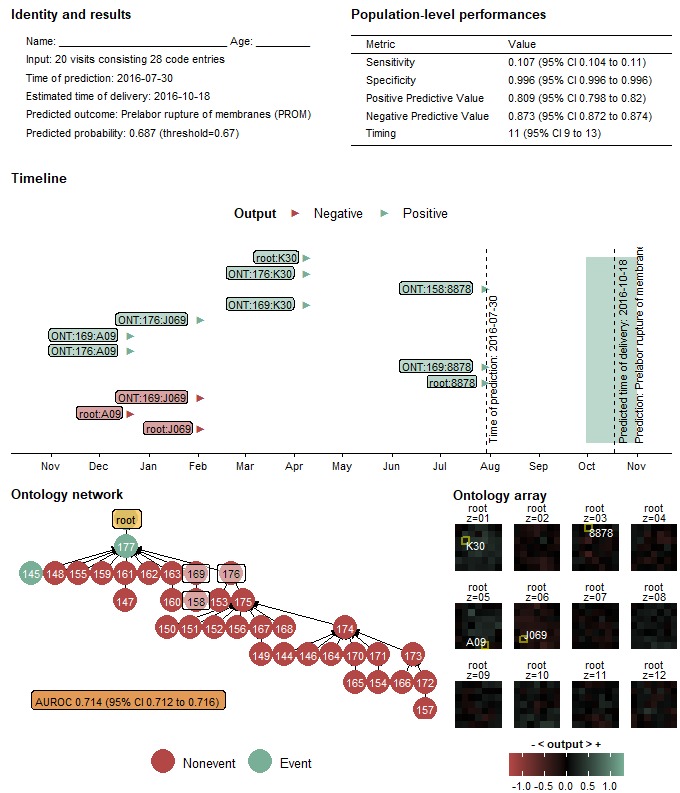
All associated factors were verified by inverse probability weighting (IPW) using our data. Association tests were only conducted between PROM and each of the associated factors. Inter-associated factor relationships were not verified, but this demonstrates how an associated factor is included in the associated model of another associated factor in this figure. APH, ante-partum hemorrhage; GTI, genital tract infection; IAI, intra-amniotic infection; PROM, prelabor rupture of membranes; SES, socio-economic status.



Exploratory data analysis

Caption:

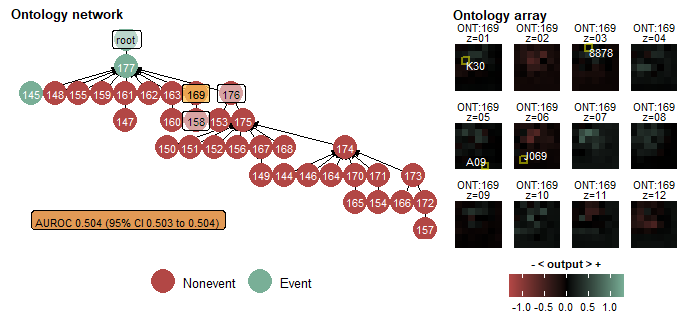
1. area under the receiver operating characteristic curve (AUROC) of DI-VNN every 4 weeks; (b) ontology (ONT) network and arrays of the DI-VNN. Showing the best time window for the prediction by the DI-VNN (a) and AUROCs of >0.55 for prediction using parts of the network architecture up to each layer on which a node resides (b). Each node is a CliXO term. We only show distinguished output arrays for a particular channel denoted by ‘z’. A feature in the array may tend to positive or negative output, color-coded based on the gradient as shown, including the feature description. Yellow squares in an array refer to a feature if only its output is non-zero. A feature may not have this square, e.g. assoc\_A03 and 8602. ONT:154 is an example of a backpropagation effect from ONT:171. CliXO, clique-extracted ontology; DI-VNN, deep-insight visible neural network.



A case example

Caption:

This is an example for predicting prelabor rupture of membranes (PROM) by DI-VNN and estimating of the time of delivery by PC-RF. An ontology term in the timeline is prefixed by ONT, followed by the number and one of the feature members. 8878, diagnostic ultrasound of gravid uterus; A09, diarrhea and gastroenteritis of presumed infectious origin; AUROC, area under receiver operating characteristics curve; DI-VNN, deep-insight visible neural network; J069, unspecified acute upper respiratory infection; K30, dyspepsia; PC, principal component; RF, random forest.



Ontology ONT:169

Association tests

Latent candidate predictors

Outcome regression (95% CI; P value)

Inverse probability weighting (95% CI; P value)

Multiple pregnancy

1.325 (1.118 to 1.569; P=.001\*\*)

1.062 (1.055 to 1.068; P<.001\*\*\*)

Chorioamnionitis

5.77 (3.823 to 8.707; P<.001\*\*\*)

1.351 (1.33 to 1.372; P<.001\*\*\*)

IAI

2.134 (1 to 4.555; P=.05)

1.118 (1.083 to 1.153; P<.001\*\*\*)

APH

0.413 (0.287 to 0.594; P<.001\*\*\*)

0.929 (0.924 to 0.933; P<.001\*\*\*)

GTI

2.138 (1.368 to 3.342; P<.001\*\*\*)

1.116 (1.101 to 1.132; P<.001\*\*\*)

Periodontal disease

0.383 (0.26 to 0.566; P<.001\*\*\*)

0.967 (0.96 to 0.973; P<.001\*\*\*)

Polyhydramnios

1.238 (0.851 to 1.801; P=.26)

0.998 (0.989 to 1.006; P>.99)

Pneumonia

0.91 (0.538 to 1.539; P=.73)

1.037 (1.025 to 1.049; P<.001\*\*\*)

Asthma

0.649 (0.514 to 0.82; P<.001\*\*\*)

0.971 (0.966 to 0.977; P<.001\*\*\*)

Low SES

0.837 (0.808 to 0.867; P<.001\*\*\*)

0.979 (0.978 to 0.98; P<.001\*\*\*)

Maternal age

0.761 (0.73 to 0.793; P<.001\*\*\*)

0.969 (0.969 to 0.97; P<.001\*\*\*)

Influenza

0.957 (0.863 to 1.061; P=.4)

0.995 (0.993 to 0.997; P<.001\*\*\*)

*P<.05;*  ***P<.01;***  P<.001; CI, confidence interval