CYO Project: Predicting pregnancy outcome by machine learning application on a nationwide health insurance dataset

Herdiantri Sufriyana, Yu-Wei Wu, Emily Chia-Yu Su

Contents

List of Figures

List of Tables

Introduction

Dataset and variables

The dataset was taken from a nationwide health insurance dataset consisting ~1.6 million patients. Although we used deidentified dataset, it's not possible to share the dataset; thus, we selected and recoded the dataset in order to hide the data source information (Table 1). For research confidentiality, we cannot report the diagnosis being the predicted outcome. Instead, we will only report the diagnosis as 'outcome'. The outcome was a particular pregnancy endpoint. We also cannot mention the references for the same reason.

Table 1: Instance selection

Selection criterion	Excluded visits	Included visits	Excluded subjects	Included subjects
Total	0	2,745,120	0	1,697,452
Age selection by 12 to 55 years old	982,688	1,762,432	494,359	1,203,093
Sex selection by female	667,546	1,094,886	610,231	592,862
Pregnancy status selection by code and description	794,000	300,886	532,107	60,755
Term selection by pregnancy visit dates	116,303	184,583	37,855	22,900
Prediction time window selection by 1 day to event at latest	70,450	114,133	531	22,369

We selected 12-to-55-years-old pregnant women with any visits to any healthcare providers nationwide. The dataset consisted of 2-year records, but we only selected any records of a patient up to 1 day before the earliest date of event visit or the latest pregnancy visit date for the non-event visit. Event is defined if a visit was done by a subject encountered with a diagnosis code of interest. If a subject is not encountered by this code, then her visits were considered as non event. Therefore, we used visits as instances (n=114,133) for developing a prediction model using predictors derived from 65 attributes.

Goal

The model predicts whether a patient will be encountered with the diagnosis code of interest. The prediction is conducted everytime she visits a healthcare provider and gets recorded into the medical database. This is why we chose up to 1 day before the event. If being predicted as event, then it may happen a day after at the earliest. Since the outcome is theoretically happening in a very specific period within pregnancy, a clinician can contextually interpret when the outcome probably happens.

Key steps

To achieve the goal of the this project, we conducted these key steps:

1. Data preprocessing (data cleaning)
2. Data exploration and visualization
3. Any insights gained
4. Modeling approach
   - Model development
   - Model calibration
   - Model evaluation
   - Model validation

Methods

Data preprocessing

First, we filtered 65 attributes that has only completed data. There were only 26 attributes that fulfill this criterion (Table 2). Briefly, the attributes covered information regarding identity codes, subject demography (age, occupation, insurance class, and marital status), healthcare provider information, diagnosis/procedure ICD-10 codes, sampling-related attributes, and outcome-related attributes.

Table 2: Non-missing attributes

Group	Attributes
1. Identity codes	est_strata_id, householder_id, subject_id, visit_id
2. Subject demography	age, est_householder, household_member, insurance_model, marital_status, occupation_segment
3. Diagnosis/procedure	icd10_3mer_desc, icd10_code, icd10_desc
3. Healthcare provider information	healthcare_city, healthcare_level, healthcare_province, healthcare_type, reghc_city, reghc_holder, reghc_province, reghc_type
4. Sampling-related attributes	sample_cat, sampling_weight
5. Outcome-related attributes	day_to_event, event_t, outcome

Then, we spread the ICD-10 codes whereas a code had a column. If a visit was encountered by a code, then value for the code column was 1; otherwise, the value was 0. Then, we computed the number of days each code was encountered for a subject up to a visit. For example, if a code is encountered twice in a day, we just record the code having value of 1. If a code is ever encountered for three days before a visit, then we record the code having value of 3 for that visit. Therefore, we count frequency of each code being encountered but only counting it once for each day. This would be medical histories for predictors. We constructed 2 types of table for medical history.

First, we count frequency of a code being encountered for each subject up to each visit. But we count the frequency without separating healthcare providers which a subject visits to. We called this table as nationwide medical history. We used this table to identify causal factors. In this table, no medical history is censored since we know the medical history of a patient across all providers.

Second, we count the frequency with separating healthcare providers which a subject visits to. We called this table as provider medical history. We used this table to develop predictions models. This reflects real-world data for deploying the models since a healthcare provider is unlikely to access medical record of a patient in other providers. Therefore, we need a prediction model that only use patient data in each provider.

We have conducted data wrangling by pivoting the dataset longer and wider multiple times. This may incidentally cause unexpected data attrition. We need to check data integrity by comparing the numbers of event and non event (Table 3).

Table 3: Data integrity

Event	Non-event	Total	Type	Dataset
6,280	107,853	114,133	Visit	Original
1,557	20,812	22,369	Subject	Original
6,280	107,853	114,133	Visit	Nationwide
1,557	20,812	22,369	Subject	Nationwide
6,280	107,853	114,133	Visit	Provider
1,557	20,812	22,369	Subject	Provider

Data partition was conducted to get internal and external validation sets for training and testing the prediction models. We also explored the data using only internal validation set. This is intended to get better generalization of the data interpretation and the prediction models. To get external validation set, we applied either geographical or temporal split, as recommended by PROBAST guidelines. We also overlapped both splits to get third partition which is geotemporal split.

For geographical split, we randomly excluded a list of cities where a healthcare provider which a subject was registered to (not a subject visit to). Unlike a visiting provider, a registering provider is only one for each subject. We tried several numbers of city proportion for each province to exclude. We excluded visits of a subject that was registered in a healthcare provider in these cities. We excluded cities for each province to avoid racial differences. Some provinces have a dominant race over others. This factor affects the outcome based on previous studies. Although this challenge may show robustness of a prediction model, race may also increase the predictive performance that may make us too optimist.

For geotemporal split, of the excluded cities, we also tried several numbers of the proportion to exclude for the second time. From this exclusion, we overlapped the visits with those from temporal split. We chose the proportions of exclusion in order to get total excluded visits as much as we need from any splitting methods. Second exclusion was needed since the geographical split tradeoff the size of the geotemporal split. Meanwhile, we need to achieve a minimum number of events for each partition of external validation set. We will explain the external validation size requirement after description of temporal splitting below.

For temporal split, we randomly exclude a period consisting a particular number of days for each season. Like racial difference, season affects the outcome based on previous studies. To avoid overoptimism, we excluded a period for each season.

After getting the list of excluded cities and event periods, we conducted data partition. First, we excluded visits from subjects registered to providers in the excluded cities. Second, we excluded visits from subjects with event date in the excluded period. For geotemporal split, geographically-excluded visits were filtered with those of secondly excluded cities and those with event in the excluded periods. For internal validation set, data were held out from any of the splits. For geographical and temporal split, we subtracted geotemporal split from each split. We also split the internal validation set applying k-fold cross validation.

Based on PROBAST guidelines, we pursued >100 event for each split of external validation (Table 4). We considered to get visits of validation split in internal validation set, being similar to those of external validation sets. Meanwhile, we need to keep >100 event; thus, several numbers of proportion were found to fulfill those criteria. For each province, we found 22.5% cities being able to exclude for geographical split, of which, 40% cities of them being able to exclude for geotemporal split. We also found 30 days for each season being able to exclude for temporal split. By ratio of 5:1:1 for training:validation:testing, we could fulfill the partition criteria. By ratio of 5:1, we could apply k-fold cross validation by k=6. This means our testing sets (external validation) comprising ~15% of our data, which approach ~20%, the rule of thumb of testing proportion that we also tried to fulfill. We expected these numbers being able to estimate the generalizability of our prediction models.

Table 4: Data partition

Set	Event	Non-event	Total	Proportion (%)
Int. validation	5,261	92,308	97,569	85.49
Int. validation, training split	4,384	76,923	81,307	71.24
Int. validation, validation split	877	15,385	16,262	14.25
Ext. validation	1,019	15,545	16,564	14.51
Ext. validation, geographical split	680	11,449	12,129	10.63
Ext. validation, temporal split	134	1,753	1,887	1.65
Ext. validation, geotemporal split	205	2,343	2,548	2.23

Data exploration and visualization

After data wrangling to get medical history, we could identify predictor candidates. There were 2,966 predictors consisting medical histories as either diagnosis or procedure ICD-10 codes, and non-medical history predictors. Of the last type, the categorical predictors were counted as binary predictor. This means a categorical predictor with 3 categories being counted as 3 predictors.

To get insight of which predictors are important, we firstly explored from previous studies to identify any factors that were correlated with the outcome. By theoretical explanation proposed by previous studies, we made causal diagram for the outcome. We merged the diagram, which is a directed acyclic graph (DAG), from all causal factors being ever proposed for the outcome (Figure 1). To make this diagram, we search for any causal factors (A) of the outcome (Y) and the confounding factors (L). We found 28 proposed causal factors with 10 confounding factors. For research confidentiality, we cannot disclose the previous studies we considered to make the causal diagram.

Based on a causal diagram of each causal factor, we conducted G-estimation to confirm the causal relationship using our data. For example, if we want to confirm A02 as the causal factor, then we used the merged diagram to get edges that are connected from and to A02 and the outcome or Y01 (Figure 2). To take measurement error into account, we applied a measurement node. A factor is measured by any of 1 or more diagnosis/procedure codes. If we did not have any data to measure a factor, then we considered there is a backdoor path we cannot adjust when conducting G-estimation for causal inference. In this example, a path of A02-A21-Y01 is considered as backdoor path since we cannot block the confounding of A02-to-Y01 causal relationship. We also insert unmeasured nodes (U) and assumed all of this measurement errors are independent non-differential. Later for the causal model in G-estimation analysis, we included A20 and A25 with A02 as covariates and Y01 as the outcome. In the end, G-estimation would determined whether A02 is causal factor, assuming the causal model is correct.

To conduct G-estimation, we need to construct dataset for causal inference. In addition to age and non-medical history categorical predictors, all of the medical history predictors were zero if these are never encountered; otherwise, this a probability from the number of days starting from 2 years before the event. The inverse probability of the earliest medical history would be an additional predictor which is censoring probability predictors. This indicated how much day proportion of 2 years before the event, that is censored (a medical history is not recorded but probably existing).

Then, we looked at each causal diagram and made a model consisting a causal factor of interest and the confounding factors, as described in the previous example. The dataset for causal inference was used but we convert all non-zero probability of medical histories into value of 1. We also used G-estimation function, as previously described.

First, we used logistic regression to conduct causal inference using the same formula. Then, we used G-estimation. If value of 1 is not covered within the 95% confidence interval odds ratio (OR), then the causal

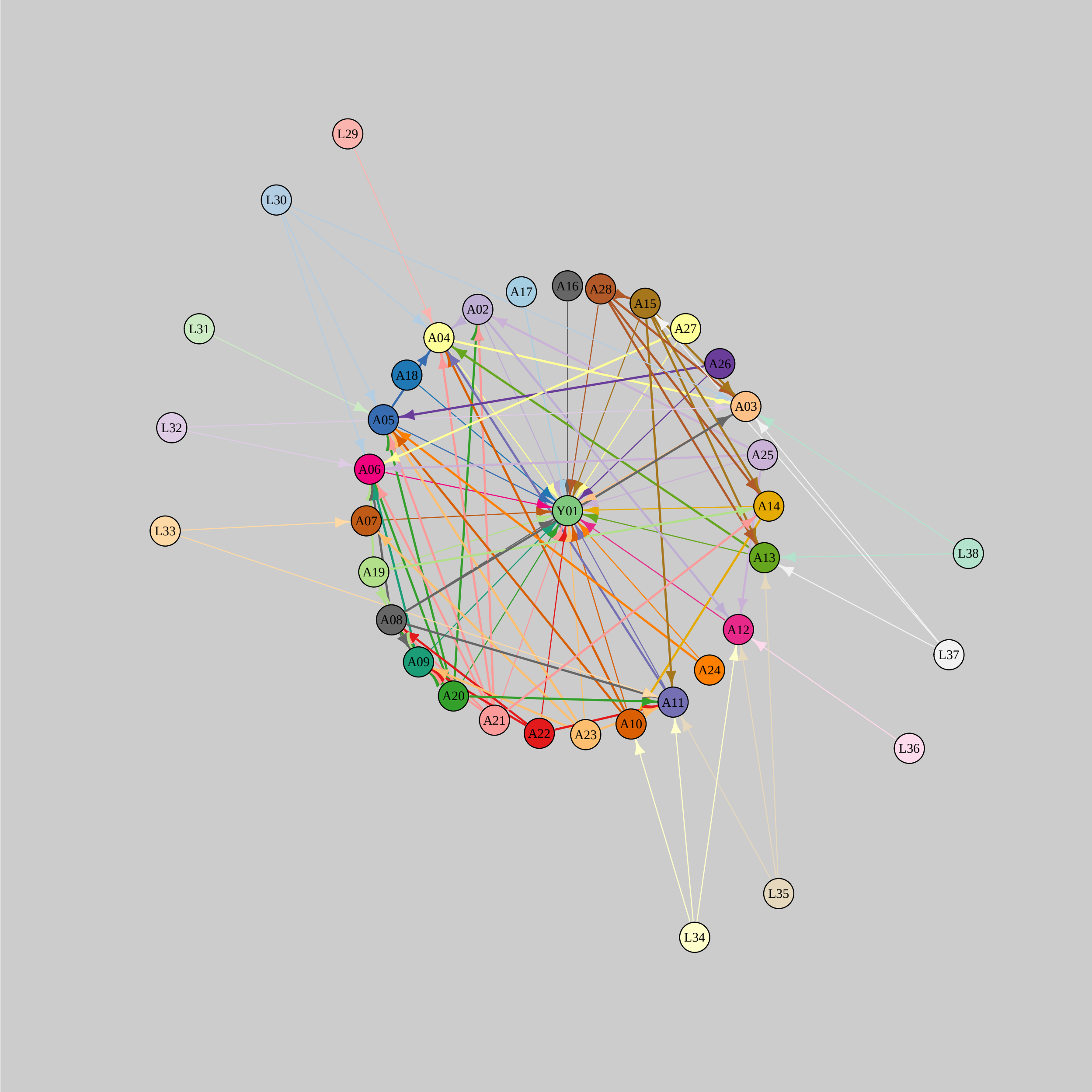

Figure 1: Merged causal diagram

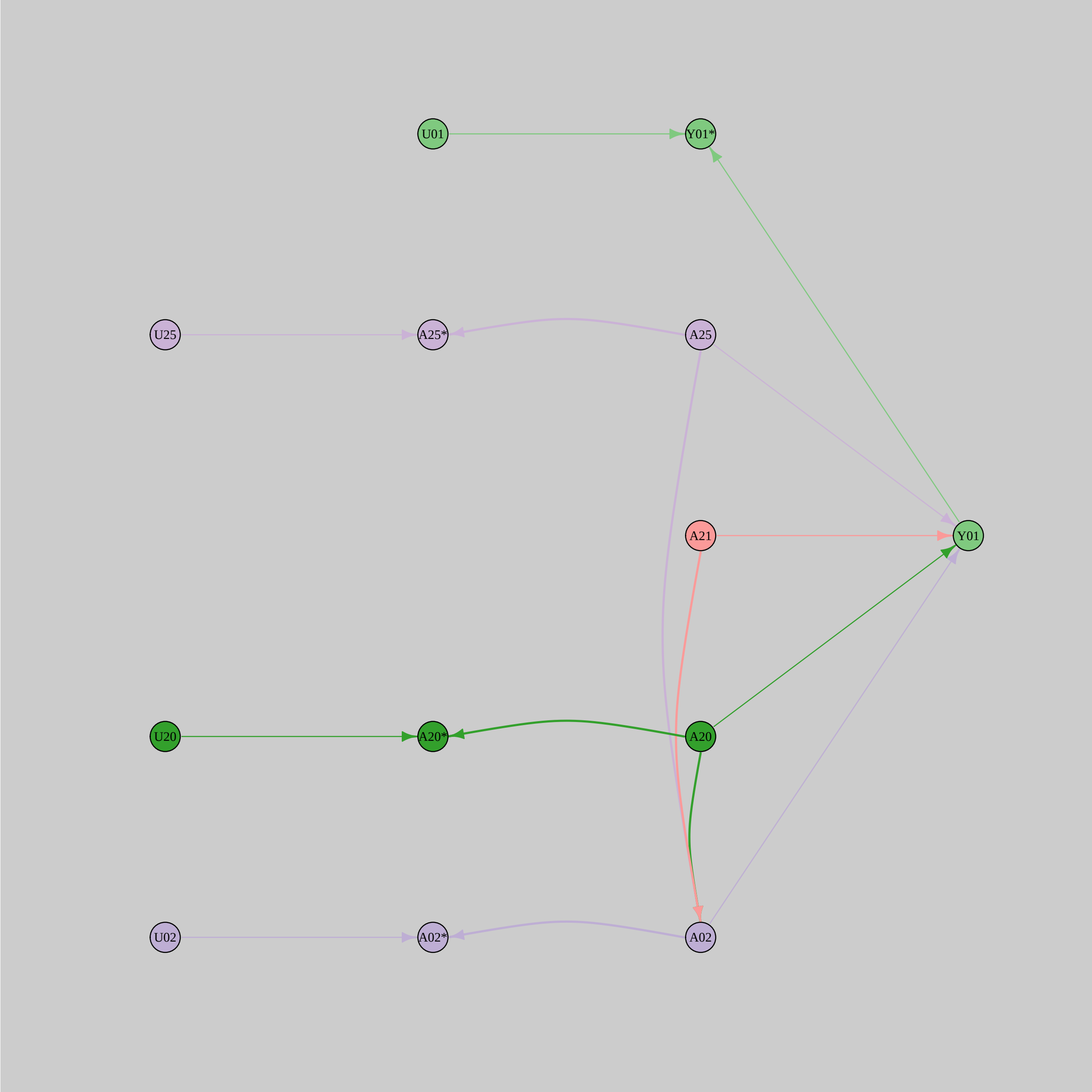

Figure 2: Example of A02 causal diagram

factor of interest is confirmed having causal relationship, assuming the causal model (diagram and formula) is correct. The results of both methods are shown (Table 5). Finally, we used results from G-estimation to determine causal factors. A previous study showed G-estimation was robust to reconstruct a model to make the simulated data for >95% of times by Monte Carlo simulation, although the causal diagram or formula used for G-estimation was incorrectly defined.

Table 5: Causal inference

Causal code	Causal factor of interest	Logistic Regression	G-estimation
A02	Multiple pregnancy	Yes	No
A03	Chorioamnionitis	No	No
A04	IAI	No	Yes
A05	Cervical shortening	No	Yes
A06	APH	No	Yes
A09	Illicit drug use	No	No
A10	GTI	No	No
A11	Periodontal disease	No	Yes
A12	Polyhydramnios	Yes	No
A13	Pneumonia	No	Yes
A14	Tuberculosis	Yes	No
A15	Asthma	No	Yes
A19	Low SES	No	Yes
A20	Maternal age	No	Yes
A24	Uterine anomaly	No	Yes
A25	Assisted reproduction	Yes	Yes
A28	Influenza	Yes	Yes

Any insights gained

If we look at the causal diagram of the confirmed causal factors (Figure 3), we could find several microbial-related factors. Other medical factors, such as assisted reproduction and uterine anomaly, were considerably differential to the microbial-related factors, and theoretically the most antecedents over other factors.

However, we realized no data being provided for other causal factor candidates (Table 6). These factors were also confounding factors in causal diagrams of other causal factors of which data were available. Except conization, none of the confounding factors is clinically relevant as causal factors. Conization is not considered as a common intervention. In addition, the prevalence of the outcome is conceivably very high in relative to conization prevalence.

Table 6: Unblocked confounding factors

Unblock causal factor candidates	Causal code
Cigarette smoking	A08
Race	A21
Low education	A22
Stress	A23
Conization	A26
Placenta on anterior wall	A27

Considering these insights, we decided to include the causal factors into predictor candidates. Of course, we still used all predictors, including any medical history, without considering whether these are confounding, since predictive modeling have a very different principle to etiologic modeling. It is also reasonable if a prediction model using only causal factors may have worse predictive performance compared to another

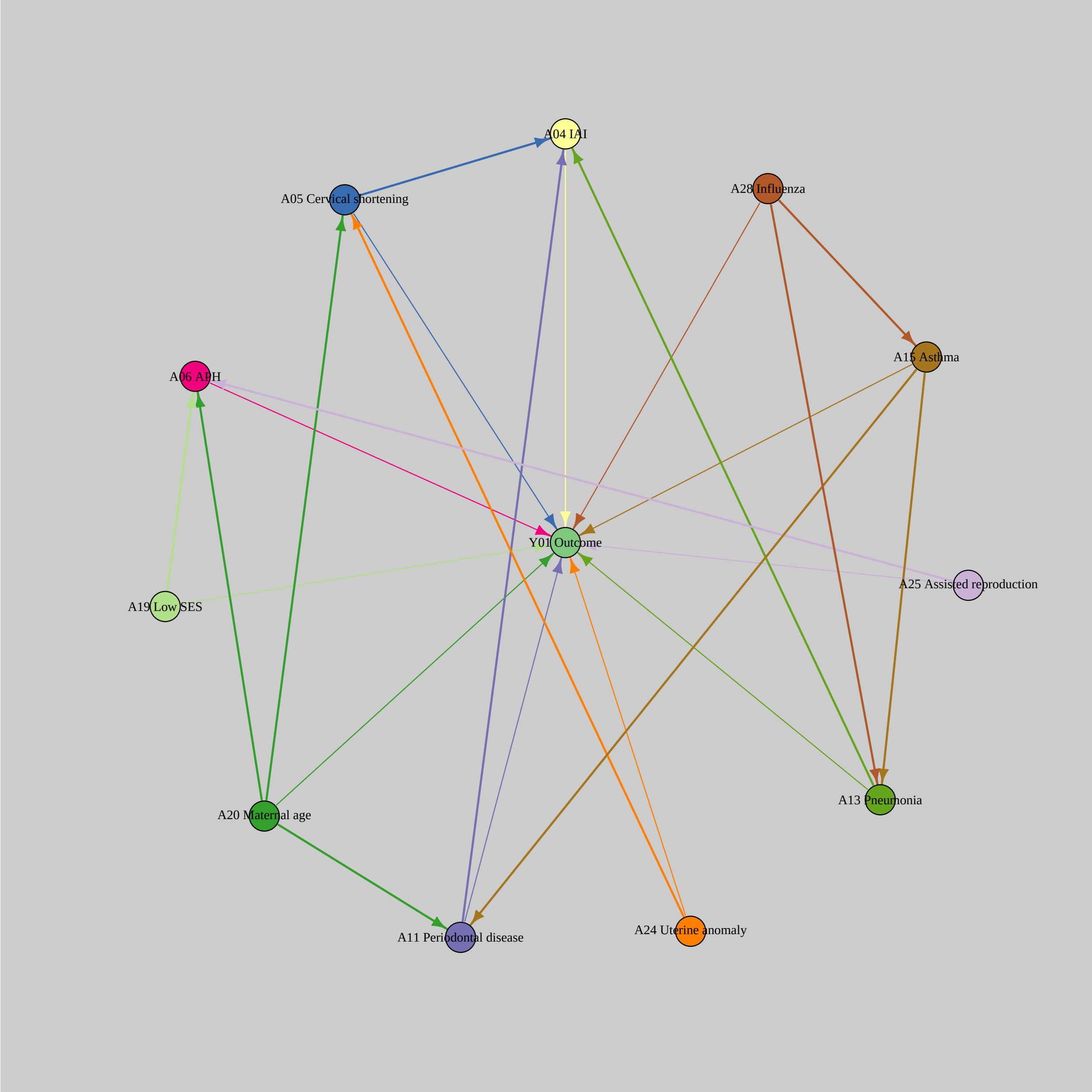

Figure 3: Final causal diagram

model using any predictors. Our approach by using causal inference is to get more insights, not only for improving predictive performance or model generalizability, but also developing a prevention strategy in future studies.

Modeling approach

We applied five approaches to develop the predictions models. One of them using only causal factors as predictors/features, while the remaining used principal components (PCs) as features. Then, we applied regression model. The selected PCs were used for last three models using different algorithms. In the end, all models would be calibrated. These are five prediction modeling approaches:

1. Ridge regression using only causal predictors (Causal LR)
2. Elastic net regression using PCs (PC-LR)
3. Random forest using selected PCs (PC-RF)
4. Linear discriminant analysis using selected PCs (PC-LDA)
5. Gradient boosting machine using selected PCs (PC-GBM)

Before developing any models, we constructed a training set using any predictors, including medical histories and causal predictors in isolated provider. Convert each of the predictors into proportions of days up to 2 years before the event, of which a predictor (a diagnosis/procedure code) being encountered. We subset only internal validation set and add a predictor of censoring probability, as previously described. In the end, we balanced the outcome by naive random oversampling to combat class imbalance problem.

Model development

Ridge regression using only causal predictors (Causal LR) We used ridge regression for causal LR because we need to retain all of the causal predictors. In this regression model, L2-norm regularization is applied as the shrinkage method. This means alpha being set to 0 while lambda was tuned using grid search from 10e-9 to 1 split into 10 numbers evenly. We applied 6-fold cross validation for determining the best tuning parameter (i.e. lambda) based on the area under receiver operating characteristics (AUROC). Using the best lambda, we trained the model and used 30-times bootstrapping to get validation AUROC interval estimates.

Elastic net regression using PCs (PC-LR) Before developing the remaining prediction models, we filter medical history predictors, including causal ones, that have non-zero variance to avoid perfect separation problem. This means one of two outcome having only one of two categories in a predictor. Although this may be true relationship between the predictor and the outcome, this situation may also happen by chance, or sampling error. Therefore, we removed this kind of predictor. We determine the variance by only using training set (internal validation).

We computed PCs using only training set by 6-fold cross validation. We standardized all predictors using average and standard deviation of training partition for each fold. This means we had 6 versions of PCs.

Then, we used top PCs that contributed cumulative percentage of variance explained for 50% at minimum (filter method). We simply determined this proportion to reduce the number of predictor candidates for further selection using PC-LR model (wrapper method).

To transform original predictors into PCs using 6 versions, we used estimates of mean, standard deviation, and the weights for each PC to transform all predictors into each PC. The estimates were computed by average. Intuition behind this calculation is that we tried to estimate PCs in population by cross validation. This procedure is similar to a linear combination of all predictors into a PC, but the weights for each PC are determined using PC analysis (PCA) pipeline. In PCA, a later PC is perpendicular to earlier one in low dimension. This is repeated until 100% variance explained and the number of maximum PCs are the same with the number of predictors that are projected. Therefore, we projected predictors into lower dimension but retaining at least 50% variance given by all predictors.

We used only PCs as features for PC-LR. We also used this model for further selection to determine feature candidates in the next models that uses other machine learning algorithms. Non-LR machine learning algorithms were shown data hungry. LR only needs at least 20 events per variable (EPV; predictor candidate), while other machine learning algorithms need up to 50 to >200 EPV. To approach the requirement, we applied this wrapper method to determine the predictor candidates for the non-LR machine learning models.

In PC-LR, we applied elastic net regression. Random search of 10 combinations of alpha and lambda was conducted. We used this shrinkage method to select important PCs by L1-norm regularization while minimizing feature exclusion by L2-norm regularization. We also included censoring probability as additional feature beyond the PCs to include information of how much data recorded for a patient. As causal LR, we applied the same validation techniques for determining either the best tuning parameter or final predictive performance using the best tuning parameter.

Since PC-LR applied L1-norm regularization, some features would be excluded; thus, we could select the features furthermore for the next prediction models. However, the exclusion were based on absolute weights arranged from the highest to the lowest, not including censoring probability. We chose top n PCs that fulfill 200 EPV. For the next models, we only used the selected PCs.

Random forest using selected PCs (PC-RF) We used random forest because it is one of two competition-winning models in machine learning. This model takes some features and some samples multiple times to construct multiple classification trees in parallel. Prediction is conducted by ensemble of the tree predictions. This deals with non-linear problem. We applied the same strategy with PC-LR for tuning parameter and final training, including the validation techniques.

Linear discriminant analysis using selected PCs (PC-LDA) We also used LDA in assumption of linearity between PCs and the outcome. Using continous independent variables from PCs, this model explicitly attempts to model the outcome difference. We also used the selected PCs based on PC-LR to get features that follows more strict assumptions of LDA. Therefore, we expected a similar model to LR, but approaching the outcome slightly different to the PC-LR. We applied the same strategy with PC-LR for tuning parameter and final training, including the validation techniques.

Gradient boosting machine using selected PCs (PC-GBM) The last model used another competition-winning model. Unlike random forest, this model constructs trees sequentially. Later tree predicts error of the earlier one. By using this model, we applied two linear models, and two non-linear models beyond the causal LR. We applied the same strategy with PC-LR for tuning parameter and final training, including the validation techniques.

Model calibration

All prediction models were calibrated by univariable logistic regression using the predicted probabilities. Before calibration, we explored the calibration plot and ROC curve. Calibration intercept and slope, and the AUROC, were also computed.

First, we computed the predicted probability using the trained models. The training set was used to compute the probabilities. We made a new training table for each model.

Then, we trained a logistic regression for each model using the predicted probability as single predictor. We applied 30-times bootstrapping on the training sets. These were the aforementioned new training tables consisting the predicted probabilities for each model.

After all prediction models were calibrated, we compared the calibration plot. We conducted the calibration to get a better probability distribution and expected a clinician having better confidence when interpreting the predicted probability. Well-calibrated models can be compared for the ROC curves furthermore. Calibration intercept and slope, and the AUROC, were also computed to get precise comparison.

Model evaluation

We chose the best model using internal validation set. Our criteria for the best model(s) is a well-calibrated model that significantly outperform AUROCs of other models by 95% confidence interval. Well-calibrated model is one with intercept and slope interval estimates covering 0 and 1, respectively. By visualizing calibration plot and probability distribution, well-calibrated model should show evenly distributed probability around the reference line. Meanwhile, a model outperforms others if the AUROC interval estimates are higher than those of others without overlapping. More than one models may be selected. Nevertheless, the best model based on external validation did not changed the best model selection by internal validation.

Model validation

For external validation, we construct testing set for each splitting method and one combining all splits. We applied the same pipeline to construct training set except we used the mean and standard deviation of age in training set to standardize age in the testing sets. We also used all versions of the PC weights inferred from the training set.

Results

Predictive performance

All models were well-calibrated after the calibration (Table 7). Although PC-RF have a perfect calibration intercept and slope, later we will find this model not showing evenly distributed probabilities for prediction. Following this model, very well-calibrated models were PC-GBM, PC-LDA, and PC-LR. Causal LR model was also well-calibrated but having higher standard deviations for both calibration intercept and slope.

Table 7: Calibration intercept and slope before calibration

Model	Calibrated	Intercept		Slope	
1. Causal LR	No	-11.94	± 6.75	25.04	± 13.77
1. Causal LR	Yes	-0.14	± 0.14	1.24	± 0.37
2. PC-LR	No	-0.01	± 0.01	1.01	± 0.02
2. PC-LR	Yes	0.00	± 0.04	1.00	± 0.08
3. PC-RF	No	-0.19	± 0.04	0.63	± 0.07
3. PC-RF	Yes	0.00	± NaN	1.00	± NaN
4. PC-LDA	No	0.07	± 0.02	0.87	± 0.04
4. PC-LDA	Yes	0.02	± 0.04	0.96	± 0.06
5. PC-GBM	No	-0.23	± 0.02	1.37	± 0.04
5. PC-GBM	Yes	-0.01	± 0.02	0.99	± 0.03

Calibration plot (Figure 4) and the probability distribution (Figure 5) are shown. PC-RF have a perfect calibration but the probability is distributed within bins of 0 to <0.1 and 0.9 to <1. This is quite counter-intuitive because there is no probabilities between 0.1 and 0.9. For causal LR, probabilities are mostly around 0.5. This is also applied to PC-LDA but having wider distribution up to ~0.1 and ~0.9. For PC-LR, the probability distribution is centered around 0.75. Unlike those linear models, PC-GBM have probability distribution with two centers at 0 and 1, similar to PC-RF, but PC-GBM have probabilities between 0.1 and 0.9. These show PC-LR, PC-LDA, PC-GBM being considered as well-calibrated based on our criteria.

All AUROCs before and after calibration are shown (Figure 6). The predictive performances were evaluated using internal validation set. PC-GBM outperforms other well-calibrated models which were PC-LDA and PC-LR. Therefore, we chose PC-GBM as the best model.

The best model achieved AUROC of 0.98 (95% CI 0.98 to 0.98) (Table 8). Because the sample size of internal validation set was very large, then we have a very narrow interval estimates and AUROC number. The AUROC of PC-RF is not exactly 1, but just being rounded to this number.

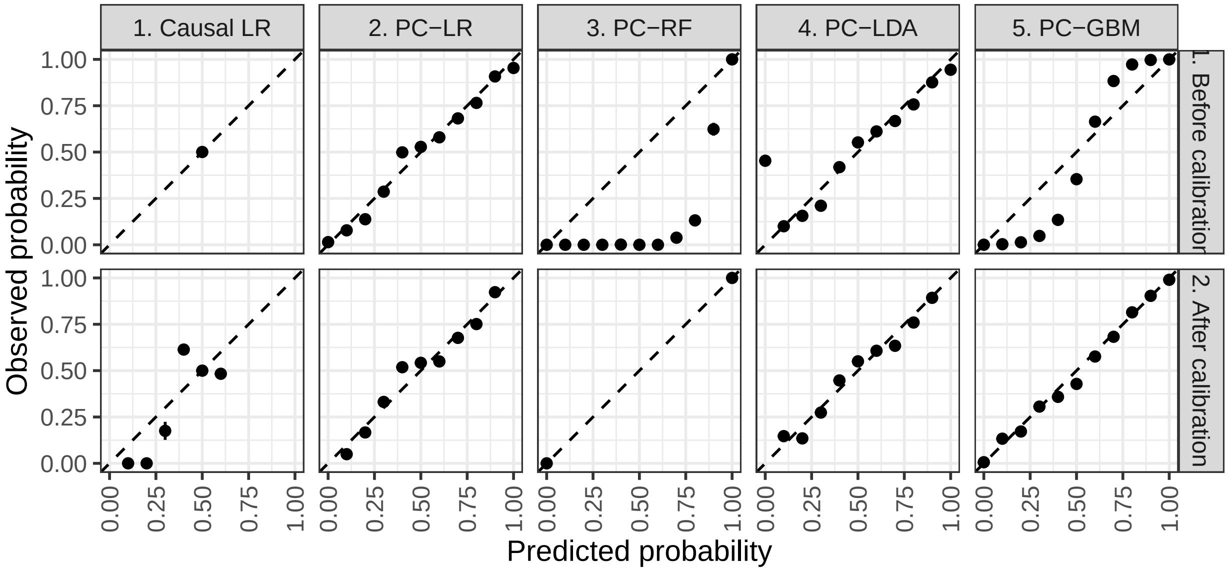

Figure 4: Calibration plot before and after calibration

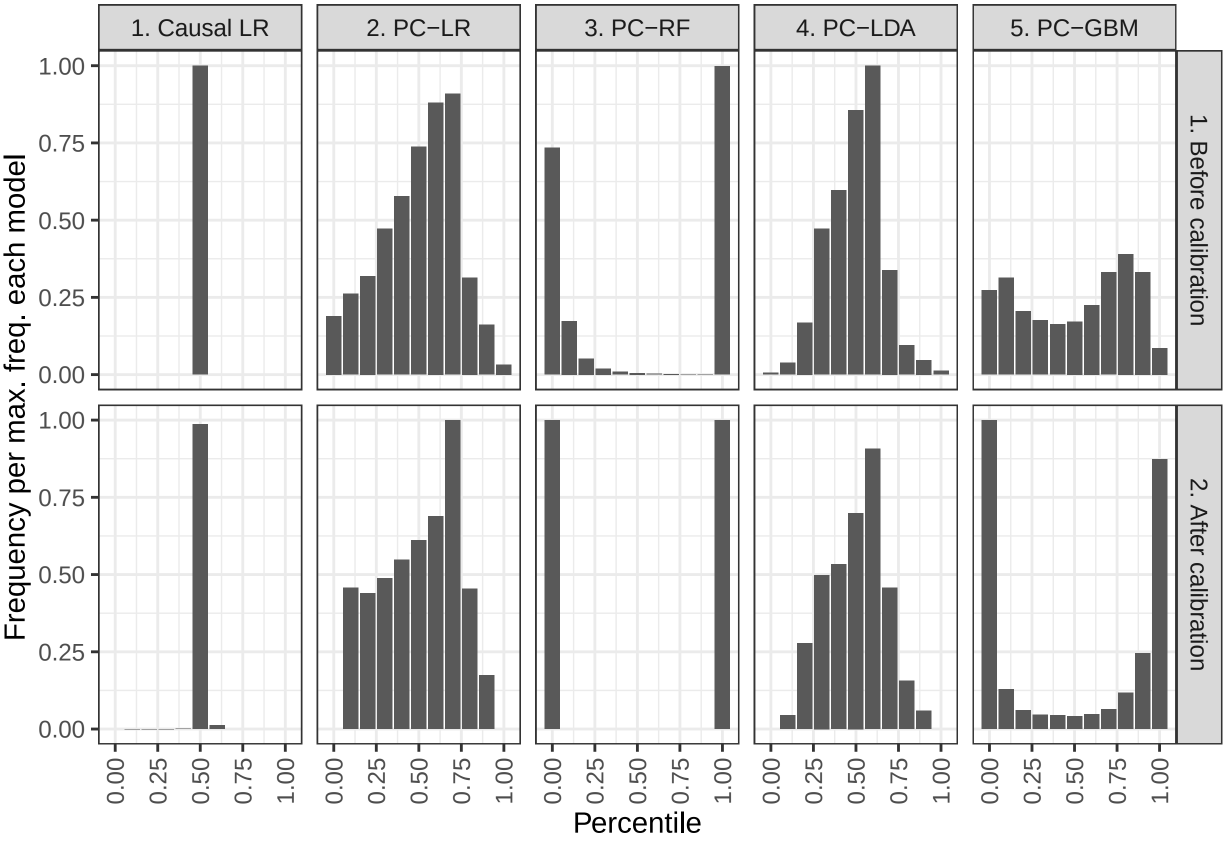

Figure 5: Probability distribution before and after calibration

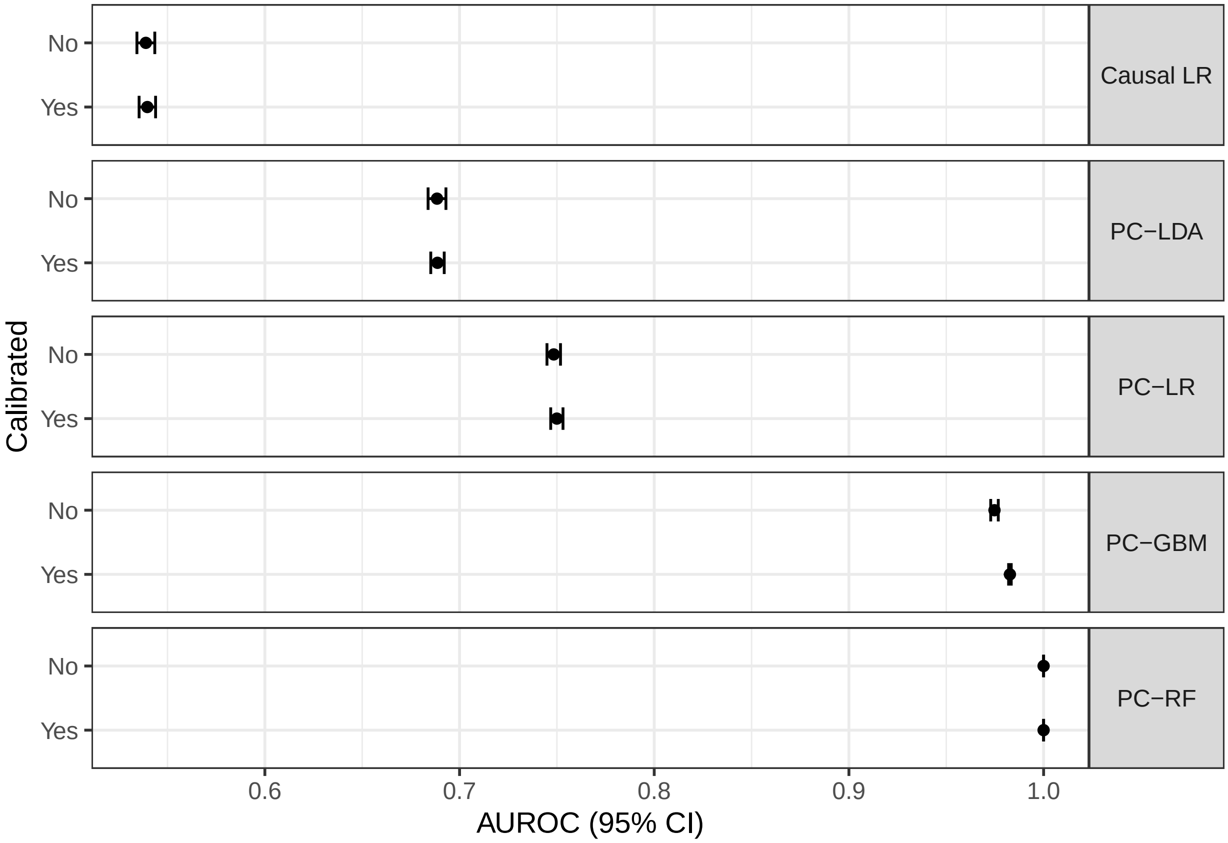

Figure 6: AUROC plot among models before and after calibration

Table 8: Compare AUROC among models before and after calibration

Model	Calibrated	AUROC	95% CI
Causal LR	No	0.54	0.53 to 0.54
Causal LR	Yes	0.54	0.54 to 0.54
PC-LDA	No	0.69	0.68 to 0.69
PC-LDA	Yes	0.69	0.69 to 0.69
PC-LR	No	0.75	0.74 to 0.75
PC-LR	Yes	0.75	0.75 to 0.75
PC-GBM	No	0.97	0.97 to 0.98
PC-GBM	Yes	0.98	0.98 to 0.98
PC-RF	No	1.00	1 to 1
PC-RF	Yes	1.00	1 to 1

After choosing the best model, we need to validate the predictive performance. We found that PC-LR is slightly better than PC-GBM in aggregated external validation set (Figure 7). This may happen by chance. We stick to choose PC-GBM as the best model to avoid overfitting. Nevertheless, of three splits for external validation sets, PC-LR was not always better than PC-GBM. Therefore, PC-GBM may still be the best prediction model for future/unobserved data.

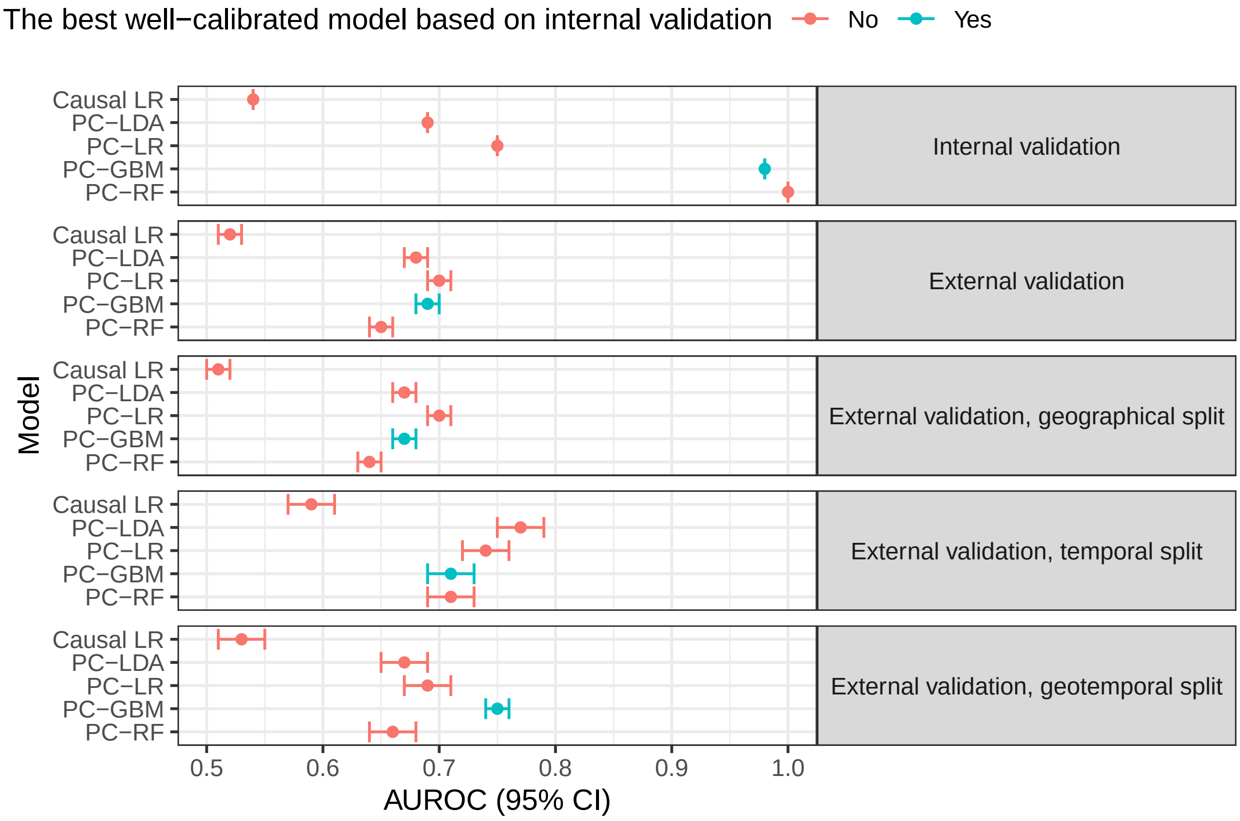

Figure 7: Plot AUROC among models

However, we found that PC-LR have a very low standard deviation of AUROCs among any sets (Figure 7 and Table 9). This reflects a very consistent predictive performance. Nevertheless, this is reasonable for linear models. Unlike these models, tree-ensemble models such RF and GBM improve predictive performances by overfitting the data. Yet, we considered these models likely to be selected using internal validation set, especially if the predictive performance is almost perfect (e.g. AUROC >0.98). This situation makes other models almost impossible to outperform RF and GBM using internal validation.

Table 9: Compare AUROC among models

Model	Validation set	AUROC	95% CI	SD among sets
Causal LR	Int. validation	0.54	0.54 to 0.54	0.031
Causal LR	Ext. validation	0.52	0.51 to 0.53	0.031
Causal LR	Ext. validation, geographical split	0.51	0.5 to 0.52	0.031
Causal LR	Ext. validation, temporal split	0.59	0.57 to 0.61	0.031
Causal LR	Ext. validation, geotemporal split	0.53	0.51 to 0.55	0.031
PC-LDA	Int. validation	0.69	0.69 to 0.69	0.042
PC-LDA	Ext. validation	0.68	0.67 to 0.69	0.042
PC-LDA	Ext. validation, geographical split	0.67	0.66 to 0.68	0.042
PC-LDA	Ext. validation, temporal split	0.77	0.75 to 0.79	0.042
PC-LDA	Ext. validation, geotemporal split	0.67	0.65 to 0.69	0.042
PC-LR	Int. validation	0.75	0.75 to 0.75	0.027
PC-LR	Ext. validation	0.70	0.69 to 0.71	0.027
PC-LR	Ext. validation, geographical split	0.70	0.69 to 0.71	0.027
PC-LR	Ext. validation, temporal split	0.74	0.72 to 0.76	0.027
PC-LR	Ext. validation, geotemporal split	0.69	0.67 to 0.71	0.027
PC-GBM	Int. validation	0.98	0.98 to 0.98	0.126
PC-GBM	Ext. validation	0.69	0.68 to 0.7	0.126
PC-GBM	Ext. validation, geographical split	0.67	0.66 to 0.68	0.126
PC-GBM	Ext. validation, temporal split	0.71	0.69 to 0.73	0.126
PC-GBM	Ext. validation, geotemporal split	0.75	0.74 to 0.76	0.126
PC-RF	Int. validation	1.00	1 to 1	0.152
PC-RF	Ext. validation	0.65	0.64 to 0.66	0.152
PC-RF	Ext. validation, geographical split	0.64	0.63 to 0.65	0.152
PC-RF	Ext. validation, temporal split	0.71	0.69 to 0.73	0.152
PC-RF	Ext. validation, geotemporal split	0.66	0.64 to 0.68	0.152

Interpretation of the best model

We used predictor weight in every PC and the PC importance in PC-GBM model to rank predictor importance. Then, we focus on predictors that were causal factors (Figure 8). Maternal age is considered the most important among these factors. This possibly because age is confounding to many predictors; thus, PC-GBM may exploit the confounding path through this factor to explain data variation for predicting the outcome.

Interestingly, by the rank, causal factors from intraamniotic infection (IAI) to pneumonia belong to a common path (Figure 3). We may consider this path as infectious/immune disease path. Meanwhile, antepartum hemorrhage (APH) and assisted reproduction belong to another common path. But, by prevalence, assisted reproduction with APH was less common compared to the conditions in the infectious/immune disease path.

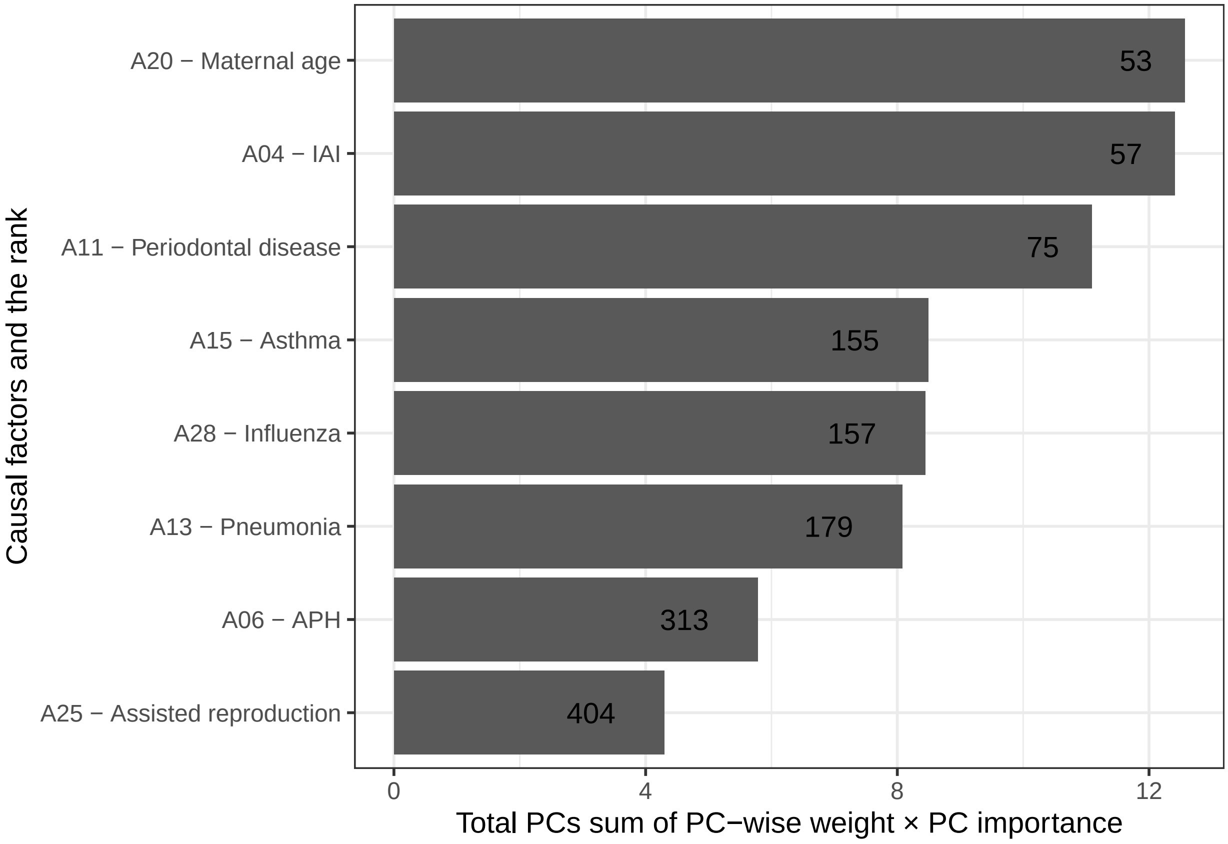

Figure 8: Causal factors by PC-GBM predictor importance

By this finding, we consider the outcome possibly caused by influenza but indirectly by modifying immune responses. These may disrupt normal microbial community within uterus. Although uterus is believed as a sterile organ for many years, recent studies show normal flora (microbial community) existing in this organ. Therefore, change in uterine normal flora may affect the outcome predicted by the PC-GBM model.

Conclusions

Brief summary

We have developed a prediction model for a pregnancy outcome. This model applied gradient boosting machine algorithm and used PCs of causal factors, medical history, and demographical variables. The model achieved 0.69 (95% CI 0.68 to 0.70) based on external validation. Influenza as the cause of the outcome was an insight gained from its rank of importance within the PC-GBM model.

Potential impact

Only one previous study have developed a prediction model for this outcome. The AUROC was 0.67 using internal validation, while our model achieved 0.98 using internal validation. This model may help improving prevention of the outcome by prediction and early intervention. Since no prevention method is already available for the outcome, insight from our prediction model may help to develop the prevention strategy, e.g. by identifying the timing based on the best one for prediction.

Limitations

Our model is still need to improve for achieving clinically-significant predictive performance using external validation. Although we can interpret the variable importance to get new insight from our prediction model, we still cannot interpret how predictor-to-predictor relationship to describe the pathogenesis model of the outcome.

We also did not have data for several causal/confounding factors. Although G-estimation is found to be robust even using an incorrectly-specified causal model, we still cannot block potential confounding since lacking of data.

Future work

Molecular studies are needed to develop both prediction and prevention for the outcome. We will use omics data to pursue these goals. A prediction model using biomarkers may improve the predictive performance. Our model in this study can be used as a preliminary prediction model to reduce healthcare cost due to the biomarker tests.

Understanding to the molecular mechanism of the outcome may help on the prevention strategy development. It should be relevant to the prediction models; thus, we may expect an integrated prediction-prevention strategy to improve this pregnancy outcome.