Dense phenotyping from electronic health records enables machine-learning-based prediction of preterm birth

Authors: Abin Abraham[1,2], Brian Le[3], Idit Kosti[3,4], Peter Straub[1,5], Digna R. Velez-Edwards[1,6,7], Lea K. Davis[1,8,9], Louis J. Muglia[10], Antonis Rokas[6,11], Cosmin A. Bejan[6], Marina Sirota[3,4], John A. Capra[1,6,11,*]

Affiliations:
[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA.
[2]Vanderbilt University Medical Center, Vanderbilt University, Nashville, TN 37232, USA.
[3]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA.
[4]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA.
[5]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.
[6]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA.
[7]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, USA.
[8]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.
[9]Department of Psychiatry and Behavioral Sciences, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.
[10]Burroughs-Wellcome Fund, Research Triangle Park, North Carolina, USA.
[11]Department of Biological Sciences, Vanderbilt University.

*Corresponding author: tony.capra@vanderbilt.edu

Abstract: Identifying pregnancies at risk for preterm birth, one of the leading causes of worldwide infant mortality, has the potential to improve prenatal care. However, we lack broadly applicable methods to accurately predict preterm birth risk. The dense longitudinal information present in electronic health records (EHRs) is enabling scalable and cost-efficient risk modeling of many diseases, but EHR resources have been largely untapped in the study of pregnancy. Here, we apply machine learning to diverse data from EHRs to predict singleton preterm birth. Leveraging a large cohort of 35,282 deliveries, we find that a prediction model based on billing codes alone can predict preterm birth at 28 weeks of gestation (ROC-AUC=0.75, PR-AUC=0.40) and outperforms a comparable model trained using known risk factors (ROC-AUC=0.59, PR-AUC=0.21). Our machine learning approach is also able to accurately predict preterm birth sub-types (spontaneous vs. indicated), mode of delivery, and recurrent preterm birth. We demonstrate the portability of our approach by showing that the prediction models maintain their accuracy on a large, independent cohort (5,978 deliveries) with only a modest decrease in performance. Interpreting the features identified by the model as most informative for risk stratification demonstrates that they capture non-linear combinations of known risk factors and patterns of care. The strong performance of our approach across multiple clinical contexts and an independent cohort highlights the potential of machine learning algorithms to improve medical care during pregnancy.

Introduction

Preterm birth, occurring before 37 weeks of completed gestation, affects approximately 10% of pregnancies globally *(1–3)* and is the leading cause of infant mortality worldwide *(4, 5)*. The causes of preterm birth are likely multifactorial since different biological pathways and environmental exposures can trigger premature labor *(6)*. Large epidemiological studies have identified many risk factors, including multiple gestations *(1)*, cervical anatomic abnormalities *(7)*, and maternal age *(8)*. Notably, even though a history of preterm birth *(9)* is one of the strongest risk factors, the recurrence rate remains low at < 30% *(10, 11)*. Additionally, maternal race influences risk for preterm birth with black women having twice the prevalence compared to white women *(1, 12)*. Preterm births have a heterogenous clinical presentation and cluster based on maternal, fetal or placental conditions *(3)*. These obstetric and systemic comorbidities (e.g. pre-existing diabetes, cardiovascular disease) can also increase risk for preterm birth *(13, 14)*.

Despite our understanding of numerous risk factors, there are no accurate methods to predict preterm birth. Some biomarkers associate with preterm birth, but their best performance is limited to a subset of all cases *(15)*. Recently, analysis of maternal cell-free RNA has emerged as a promising approach *(16)*, but initial results were based on a small pregnancy cohort and require further validation. In silico classifiers based on demographic and clinical risk factors have the advantage of not requiring serology or invasive testing. However, even in large cohorts (>1 million individuals), demography- and risk-factor-based models report poor to moderate performance for clinical application *(17–21)*. To date, we lack effective screening tools and preventative strategies for prematurity *(22)*.

Electronic health records (EHRs) are scalable, readily available, and cost-efficient for disease-risk modeling *(23)*. EHRs capture longitudinal data across a broad set of phenotypes with high temporal resolution. EHR data can be combined with socio-demographic factors and family medical history to comprehensively model disease risk. EHRs are also increasingly being

augmented by linking patient records to molecular data, such as DNA and laboratory test results. Since preterm birth has a substantial heritable risk *(24)*, combining rich phenotypes with genetic risk may lead to better prediction.

Machine learning models have shown promise for accurate risk stratification across a variety of clinical domains *(25–27)*. However, despite the rapid adoption of machine learning in translational research, a review of 107 risk prediction studies reported that most models used only few variables, did not consider longitudinal data, and rarely evaluated model performance across multiple sites *(28)*. Some medical domains have yet to incorporate machine learning methods. Pregnancy research is especially well poised to benefit from machine learning approaches *(29)*. Per standard of care during pregnancy, women are carefully monitored with frequent prenatal visits, medical imaging, and clinical laboratories tests. Compared to other clinical contexts, pregnancy and the corresponding clinical surveillance occur in a defined timeframe based on gestational length. Thus, EHRs are well-suited for modeling pregnancy complications, especially when combined with the well documented outcomes at the end of pregnancy.

In this study, we combine multiple sources of data from EHRs to predict preterm birth using machine learning. From Vanderbilt's EHR database (≥3.2 Million records) and linked genetic biobank (≥100,000 individuals), we identified a large cohort of women (n=35,282) with documented deliveries at Vanderbilt. We trained models (gradient boosted decision trees) that combine demographic factors, clinical history, laboratory tests, and genetic risk with billing codes (ICD-9 and CPT) to predict preterm birth. We find models trained on all billing codes from the mother's EHR distinguished preterm births from term and postterm births with high accuracy compared to other EHR features. We assess the clinical potential of these models by quantifying performance across different contexts. When restricting features to those available at different stages in pregnancy, billing-code-based models can accurately predict preterm birth at 28 weeks of gestation. Furthermore, this approach maintains high accuracy for predicting spontaneous preterm birth and preterm risk among mothers with a history of preterm birth. Finally, we demonstrate the generalizability of this approach by evaluating billing-code-based models on an external, independent cohort from University of California, San Francisco (UCSF, n=5,978). Prediction models trained at Vanderbilt maintain high accuracy in the external cohort with only a modest drop in performance. Our findings provide a proof-of-concept that machine learning on rich phenotypes in EHRs show promise for portable, accurate, and non-invasive prediction of preterm birth. The strong predictive performance across clinical context and preterm birth subtypes argues that machine learning models have the potential to add value during the clinical management of pregnancy.

Results

Characteristics and phenotyping of delivery cohort from Vanderbilt EHRs

From the Vanderbilt EHR database (~3.2 Million patients), we identified a 'delivery cohort' of 35,282 women with at least one delivery in the Vanderbilt hospital system (Fig. 1A). In addition to ICD and CPT billing codes, we extracted demographic data, past medical histories, obstetric notes, clinical labs, and genome-wide genetic data for the delivery cohort when available. Because billing codes were the most prevalent data in this cohort (n=35,282), we quantified the pairwise overlap between billing codes and each other data type. The largest subset included

women with billing codes paired with demographic data (n=33,570). The smallest subset was women with billing codes paired with genetic data (n=905; Fig. 1C). The mean maternal age of women at the first delivery in the delivery cohort was 27.3 years (Fig. 1D). The majority of women in the cohort self- or third-party reported as white (n=21,343), black (n=6,178), or Hispanic (n=3,979; Fig. 1E). The estimated gestational age (EGA) distribution had a mean of 38.5 weeks (38.0 to 40.3 weeks, 25th to 75th percentile; Fig. 1F). The rate of multiple gestations (e.g. twins, triplets) was (7.6%, n=1,353). Since multiple gestation pregnancies are more likely to deliver preterm, we developed prediction models using singleton pregnancies unless otherwise stated.

We used billing codes and EGA to ascertain the delivery date and type (preterm vs. not-preterm, Methods). In the delivery cohort, we identified 7,774 preterm births. To evaluate the accuracy of ascertaining preterm births, a domain expert blinded to the delivery type reviewed clinical notes from 104 EHRs selected at random from the delivery cohort. The ascertainment algorithm had precision/positive predictive value (PPV) of 96% and recall/sensitivity of 96% using the chart reviewed label as the gold standard (Fig. S1A).

Boosted decision trees using billing codes identify preterm deliveries

Using this richly phenotyped delivery cohort, we evaluated how well the clinical phenome, defined as only billing codes (ICD-9 and CPT) before and after delivery, could identify preterm births. With counts of each billing code (excluding those used to ascertain delivery type), we trained gradient boosted decision trees *(30)* to classify each mother's first delivery as preterm or not-preterm (Fig. 2A). Boosted decisions trees are well-suited for EHR data because they require minimal transformation of the raw data, are robust to correlated features, and capture non-linear relationships *(31)*. Moreover, boosted decision trees have been successfully applied on a variety of clinical tasks *(32–34)*.

In all evaluations, we held out 20% of the cohort for testing and used the remaining 80% for training and validation (Fig. 2A). Boosted decision tree models trained on ICD-9 and CPT codes accurately identified preterm births (singletons and multiple gestations) with PR-AUC=0.86 (chance=0.22) and ROC-AUC=0.95 (Fig. S2 A and B). While the combined ICD-9 and CPT based model achieved the best performance, models trained on either ICD-9 or CPT individually also performed well (PR-AUC ≥0.82; chance=0.22, ROC-AUC ≥0.93). All three models demonstrated good calibration with low Brier scores (≤0.092; Fig. S2C). Thus, billing codes across an EHR show potential as a discriminatory feature for predicting preterm birth.

Accurate prediction of preterm birth at 28 weeks of gestation

To evaluate preterm birth prediction in a clinical context, we trained a boosted decision tree model (Fig. 2A) on billing codes present before each of the following timepoints: 0, 13, 28, and 35 weeks of gestation (Fig. 2B). We downsampled to achieve comparable number of singleton deliveries across each timepoint to mitigate sample size as a potential confounder while comparing performance. We only considered active pregnancies at each timepoint; for example, a delivery at 29 weeks would not be included in the 35 week model, since the outcome would already be known. The ROC-AUC increased from conception (0 weeks; 0.63) to the highest performance at 35 weeks (0.75; Fig. 2C). The PR-AUC (Fig. 2D), which accounts for preterm

birth prevalence, obtains the strongest performance at 28 weeks (0.33, chance=0.13). However, as we show in the next section, this is an underestimate of the ability to predict preterm delivery at 28 weeks due to our downsampling of the number of training examples. As expected, when we included multiple gestations, the model performed even better (PR-AUC=0.42 at 28 weeks, chance=0.14; Fig. S3). Results were similar when models were trained using billing codes available before different timepoints from the date of delivery (Fig. S4).

To confirm that the number of contacts with the health system was not driving performance, we trained a classifier based on the total number of codes in an individual's EHR before delivery to predict preterm birth. This simple classifier failed to discriminate between delivery types with PR-AUC and ROC-AUC only slightly higher than chance (PR-AUC=0.19, chance=0.19; ROC-AUC=0.56, chance=0.5, Fig. S5). Therefore, cumulative disease burden or the number of contacts alone are not informative in predicting preterm birth.

Integrating other EHR features does not improve model performance

In addition to billing codes, EHRs capture aspects of an individual's health through different types of structured and unstructured data. We tested whether incorporating additional features from EHRs can improve preterm birth prediction. Models were evaluated using data available at 28 weeks of gestation, given that it is sufficiently early for intervention and enabled accurate predictions using billing codes. From the EHRs, we extracted sets of features including demographic variables (age, race), clinical keywords from obstetric notes, clinical lab tests ran during the pregnancy, and predicted genetic risk (polygenic risk score for preterm birth). To measure the performance gain for each feature set, we compared models trained using: the feature set only, billing codes only, and billing codes combined with the feature set (Fig. 3A). Within each feature set, the same pregnancies comprised the training and held-out sets for the three models. However, the number of deliveries (training + held-out sets) varied widely across feature sets (n=20,342 to 462) due to the differing availability of each feature type.

Models using only demographic factors, clinical keywords, and genetic risk had ROC-AUC and PR-AUC similar to chance (Fig. 3B). Clinical labs had moderate predictive power with ROC-AUC of 0.63 and PR-AUC of 0.24 (Fig. 3B). Compared to models using only billing codes, adding additional feature sets did not substantially improve performance (Fig. 3B). We note that some features sets, such as clinical labs and genetic risk, were evaluated on held-out sets with small numbers of deliveries (180 and 92, respectively). However, even after increasing the sample size by including EHR features present before and after delivery, we did not observe a consistent gain in performance compared to models trained using only billing codes (Fig. S6).

Models using billing codes outperforms prediction from risk factors

Although there are well known risk factors for preterm birth, there exists no clinical risk calculator that is routinely implemented in clinical care. We sought to compare the performance of our EHR-based prediction models to known risk factors. Such risk factors would inform a physician's gestalt for risk-stratifying a pregnancy. We compared the 28 week billing-code-based model to a model trained using a set of known clinical risk factors *(17)* that included: self- or third-party reported race (Black, Asian, or Hispanic), age at delivery (> 34 or <18 years old),

diabetes status, sickle cell disease status, presence of fetal abnormalities, pre-pregnancy BMI >35, and pre-pregnancy hypertension (blood pressure > 120/80, Methods).

The billing-code-based model significantly outperformed a model trained with clinical risk factors at predicting preterm birth at 28 weeks of gestation (PR-AUC=0.40 vs. 0.21; Fig. 4B). The pattern was similar for ROC-AUC (risk factors=0.59, billing codes=0.75; Fig. S7). The stronger performance of the billing-code-based classifier was true for women across the spectrum of comorbidity burden. It had higher precision across individuals with different numbers of risk factors. Performance peaked for individuals with 0 (precision=0.39) and 4+ (precision=0.43) risk factors, but we did not observe a trend between model performance and increasing number of clinical risk factors (Fig. 4C). This suggests that combinations of billing codes have the potential to quantify preterm birth risk better than risk factors that are currently used to inform clinical judgement.

Machine learning models can predict spontaneous preterm births

The multifactorial etiologies of preterm birth lead to clinical presentations with different comorbidities and trajectories. Medically-indicated and idiopathic spontaneous preterm births are distinct in etiologies and outcomes. Identifying pregnancies that ultimately result in spontaneous preterm deliveries is particularly valuable, and we anticipated that spontaneous preterm birth would be more challenging to predict than preterm birth overall. To test this, we identified spontaneous preterm births in the held-out set (n=75) at 28 weeks of gestation by excluding women with medically induced labor, a cesarean section delivery, or PPROM (Methods). We intentionally used a conservative phenotyping strategy that aimed to minimize false positive spontaneous preterm births to evaluate the model's ability to predict spontaneous preterm births. The prediction model trained using billing codes up to 28 weeks of gestation classified 48% (recall) of all spontaneous preterm births as preterm; this is significantly higher than the risk factor only model (recall = 33%; Fig. 4D).

Performance varies based on clinical context and delivery history

To further explore the sensitivity of the performance of our approach to clinical context and patient history, we evaluated how delivery type (vaginal vs. cesarean-section) and a previous preterm birth influence preterm birth prediction. We trained two classifiers using billing codes (ICD-9 and CPT) occurring before 28 weeks of gestation: one on a cohort of cesarean-section (n=5,475) singleton deliveries and one on vaginal deliveries (n=15,487). Preterm birth prediction accuracy was higher in the cesarean-section cohort (PR-AUC=0.47, chance=0.20) compared to the vaginal delivery cohort (PR-AUC=0.23, chance = 0.10; Fig. 5A). Cesarean-sections also had higher ROC-AUC compared to vaginal deliveries (0.75 vs. 0.68, Fig. S8). As expected, the preterm birth prevalence was higher in the cesarean-section cohort.

Women with a history of preterm birth are at significantly higher risk for a subsequent preterm birth than women without a previous history. Therefore, we tested if models trained on EHR data of women with a history of preterm birth could accurately predict the status of their next birth. We assembled 1,416 women with a preterm birth and a subsequent delivery in the cohort and split them into a training set (80%) and held-out set (20%) to evaluate the model performance (Methods). For these women, 53% of the second deliveries were preterm. Due to

limited availability of estimated gestational age data for the recurrent preterm births, which is necessary to approximate the date of conception, we trained models using billing codes (ICD-9 and CPT) present before each of the following timepoints: 10, 30, and 60 days before the delivery. These models were all able to discriminate term from preterm deliveries better than chance (Fig. 5B; PR-AUCs≥0.75). The model predicting a second preterm birth at 10 days before delivery achieved the highest performance with PR-AUC=0.84 (Fig. 5B, chance=0.53) and ROC-AUC=0.82 (Fig. S9).

Models accurately predict preterm birth in an independent cohort

To evaluate whether preterm birth prediction models trained on the Vanderbilt cohort performed well on EHR data from other databases, we compared their performance on the held-out Vanderbilt cohort (n=4,215) and an independent cohort from UCSF (n=5,978). The UCSF cohort was ascertained using similar rules as the Vanderbilt cohort (Methods); age and distribution of race are provided in Table S1. However, we note that the UCSF cohort has a lower preterm birth prevalence (6%) compared to the Vanderbilt cohort (13%).

To facilitate the comparison, we trained models to predict preterm birth in the Vanderbilt cohort using only ICD-9 codes present before 28 weeks of gestation. We did not consider CPT codes in this analysis due to differences in the available billing code data between Vanderbilt and UCSF. As expected from the previous results, the model accurately predicted preterm birth in the held-out set from Vanderbilt (PR-AUC of 0.34, chance=0.12), but performance was slightly lower than using ICD and CTP codes (Fig. 4B).

The model trained at Vanderbilt also achieved strong performance in the UCSF cohort (PR-AUC of 0.31 vs 0.34 at Vanderbilt; Fig. 6A). The classifier had a higher ROC-AUC (0.80) in UCSF cohort compared to the Vanderbilt cohort (0.72; Fig. S10). This is likely due to the lower prevalence of preterm birth in the UCSF and the sensitivity of ROC-AUC to class imbalance *(35)*. Overall, these models show striking reproducibility across two independent cohorts.

Similar features are predictive across the independent cohorts

The architecture of boosted decision trees enables straightforward identification of features (ICD-9 codes) with the largest impact on the model predictions. We used SHapley Additive exPlanation values (SHAP) *(36, 37)* to quantify the marginal additive contribution of each feature to the model predictions for each individual. For each feature in the ICD-9-based model, we calculated the mean absolute SHAP values across all women in the held-out set. The mean absolute SHAP value for each feature was highly correlated (spearman R=0.93, p-value < 2.2E-308) between the held-out Vanderbilt set and the UCSF cohort (Fig. 6B). Ten of the top 15 features ranked based on the mean absolute SHAP value were shared across both cohorts. Examination of the codes driving prediction revealed many known risk factors such as fetal abnormalities, history of twin pregnancy, history of preterm birth, diabetes, and other comorbidities (Fig. 6C). The majority of the top features involved codes indicating screening, routine or otherwise, during pregnancy. Three top features only in the UCSF dataset included codes for supervision of high-risk pregnancies (Fig. 6C).

Discussion

Preterm birth is a major health challenge affecting 5-20% of pregnancies *(1, 2, 12)* and leading to significant morbidity and mortality *(38, 39)*. Predicting preterm birth risk could inform clinical management, but no accurate classification strategies are routinely implemented *(22)*. Here, we take a step toward addressing this need by demonstrating the potential for machine learning on dense phenotyping from EHRs to predict preterm birth. Our models predict preterm birth accurately across challenging clinical contexts (e.g., spontaneous and recurrent) at 28 weeks of gestation. Compared to other data types in the EHRs, models using billing codes alone had the highest prediction accuracy and outperformed those using clinical risk factors. Demonstrating the potential broad applicability of our approach, the model accuracy remained high in an external independent cohort. Combinations of many known risk factors and patterns of care drove prediction; this suggests that the algorithm builds on existing knowledge. Thus, we conclude that machine learning based on EHR data has the potential to predict preterm birth accurately across multiple healthcare systems.

Our models have several distinct advantages compared to published approaches. First, they have robust performance. Previous models using risk factors (diabetes, hypertension, sickle cell disease, history of preterm birth) to predict preterm birth, despite having cohorts up to two million women *(17)*, have reported ROC-AUCs between 0.69 and 0.74 *(18, 19, 21)*. Our models obtain a ROC-AUC of 0.75 and PR-AUC of 0.40 using data available at 28 weeks gestation. Furthermore, given the unbalanced classification problem (preterm births are less common than non-preterm), we report high PR-AUCs in addition to high ROC-AUCs. Compared to a recent deep learning model using word embeddings from EHRs for predicting extreme preterm birth (birth before 28 weeks of gestation, ROC-AUC of 0.83, *40)*, our models achieved similar accuracy using only billing codes to predict all preterm births. We did not stratify preterm births by severity since more than 85% of preterm births occur after 32 weeks of gestation *(56)*. However, this is an interesting topic for further work.

Second, our models use readily available data throughout pregnancy that do not require invasive sampling. While some studies have also obtained high ROC-AUCs (e.g., 0.81-0.88), they used serum biomarkers across small cohorts *(16)* or acute obstetric changes within days of delivery *(20)*. This can enable cost-effective and broad application as illustrated by our evaluation of the classifiers on EHR data from UCSF.

Third, the gradient boosted decision trees we implement are more interpretable than 'black-box' deep learning models that cannot easily identify features driving predictions. This ability could lead to better understanding of the risk factors and differences in risk factors in different regions of the country or the world. The ease of interpretation of our decision trees is a necessary factor for future deployment in clinical settings. Our models rediscovered several known risk factors for preterm birth, which establishes further confidence in our machine-learning-based risk prediction models.

Finally, our approach generalizes across hospital systems. We demonstrate that billing-code-based models trained at Vanderbilt achieve similar accuracy in an independent cohort from UCSF. The generalizability of machine learning models can be constrained by the sampling of the training data. Thus, the accurate prediction in an independent dataset from an external institution points to several inherent strengths of the model. First, successful replication indicates the models' ability to learn predictive signals despite regional variation in assigning billing codes

to an EHR. Second, the large cohorts used to train and evaluate models at Vanderbilt and USCF guard against potential weakness of EHRs. Miscoding or omission of key data points are unavoidable in EHRs *(41)*. The large cohort used to train our models mitigates these errors and enables the high accuracy in the UCSF dataset, even with its different demographics. Additionally, idiosyncratic patterns of patient care at the institution used to develop the algorithm, which would be present in the Vanderbilt training and held-out sets, are unlikely to be present in the external UCSF cohort and inflate the out-of-sample accuracy. Third, the top features driving model performance are shared across institutions and reflect combinations of known risk factors and patterns of care. This aids interpretability of the underlying algorithm and likely reflects underlying pathophysiology that is innate to preterm birth.

We see several avenues for further improving our algorithm. First, some of the top features reflected routine obstetric care for high-risk pregnancies. Thus, the learning problem could be engineered to force the algorithm to discover new unappreciated risk factors. Second, we were surprised that the addition of features beyond billing codes, such as lab values, concepts extracted from clinical notes, and genetic information did not significantly improve performance. In some cases, any redundant information already captured by the billing codes would not improve the model's accuracy; this is likely true for clinical notes. However, other sources, like currently available genetic data and polygenic risk scores, may not effectively capture underlying etiologies of preterm birth; thus, these sources may not add more discriminatory power. Indeed, the largest published genome-wide study for preterm birth only explains a very small fraction of the heritability *(24)*, and a polygenic risk score derived from it was not predictive in our cohort. Further sub-phenotyping of preterm birth will not only aid in prediction, but also understanding its multifactorial etiology and developing personalized treatment strategies. More explicit modeling of the temporal dependence between EHR features may further increase performance. Finally, while we evaluated the ability of our classifiers to discriminate preterm births, further studies evaluating the calibration of these models are necessary to better risk stratify of pregnancies.

The strong predictive performance of our models suggests that they have the potential to be clinically useful. Compared to a machine learning model trained using only known risk factors, the billing-code-based classifier incorporated a broad set of clinical features and predicted preterm birth with higher accuracy. Furthermore, the superior performance was not driven by the number of risk factors or the total burden of billing codes. These results indicate the algorithm is not simply identifying less healthy individuals or those with greater healthcare usage. The models also accurately predicted many preterm births in challenging and important clinical contexts such as spontaneous and recurrent preterm birth. Spontaneous preterm births are common *(1, 12, 42)*, and unlike iatrogenic deliveries, they are more difficult to predict because they are driven by unknown multifactorial etiologies*(12, 22)*. Similarly, since a prior history of preterm birth is one of the strongest risk factors *(43)*, distinguishing pregnancies most at risk for recurrent preterm birth has potential to provide clinical value.

However, additional work is needed before this approach is ready for clinical application. Though it has strong performance, a more comprehensive evaluation of the algorithm against current clinical practice is needed to determine how early and how much improvement in standard of care this approach could provide *(44)*. Furthermore, while our cohorts include diverse individuals and the algorithm generalizes well, the approach must be evaluated to ensure that it does not introduce of amplify biases against specific groups or types of preterm birth *(45)*.

In addition, we anticipate further gains in the clinical value of this approach as more modalities of data becomes incorporated in the EHR *(46)* and diverse populations become available. Addressing these questions and taking other necessary steps toward clinical utility will require the close collaboration of diverse experts from basic, clinical, social, and implementation sciences.

Our results provide a proof-of-concept that machine learning algorithms can use the dense phenotype information collected during pregnancy in EHRs to predict preterm birth. The significant prediction accuracy across clinical contexts and compared to existing risk factors suggests such modeling strategies can be clinically useful. We are optimistic that with the ever-growing number of EHRs, improvement in tools for extracting meaningful data from them, and integration of complementary molecular data, machine learning approaches can improve the clinical management of preterm birth.

Materials and Methods

Ascertaining delivery type and date for Vanderbilt cohort

We identified women with at least one delivery (n=35,282, 'delivery-cohort') at Vanderbilt Hospital based on the presence of delivery-specific billing codes (ICD-9/10 and CPT) or estimated gestational age (EGA) documented in the EHR. Combining delivery specific ICD-9/10 ('delivery-ICDs'), CPT ('delivery-CPTs'), and EGA data, we developed an algorithm to label each delivery as preterm or not preterm. Women with multiple gestations (e.g. twins, triplets) were identified using ICD and CPT codes and exclude for singleton-based analyses. See Supplementary Materials and Methods for exact codes.

We demarcate multiple deliveries by grouping delivery-ICDs in intervals of 37 weeks starting with the most recent delivery-ICD. This step is repeated until all delivery-ICDs in a patient's EHR are assigned to a pregnancy. We chose 37-week intervals to maximally discriminate between pregnancies. For each delivery, we assign a list of labels (preterm, term, or postterm) ascertained using the delivery-ICDs. EGA values were mapped to multiple pregnancies using the same procedure. The most recent EGA value determined the time interval to group preceding EGA values. Based on the most recent EGA value for each pregnancy, we assigned labels to each delivery (EGA <37 weeks: preterm; ≥37 and <42 weeks: term, ≥42 weeks: postterm). After pooling delivery labels based on delivery-ICDs and EGA, we assigned a consensus delivery label by selecting the oldest gestational age based classification (i.e. postterm > term > preterm).

Since CPT codes do not encode delivery type, we combined the delivery-CPTs with timestamps of delivery-ICDs and EGAs to approximate the date of delivery. Delivery-CPTs were grouped into multiple pregnancies as described above. The most recent timestamp from delivery-CPTs, delivery-ICDs, and EGA values was used as the approximate delivery date for a given pregnancy.

Validating delivery type based on chart review

To validate the delivery type ascertained from billing codes and EGA, we used chart-reviewed labels as the gold standard. For 104 randomly selected EHRs from the delivery cohort, we

extracted the date and gestational age at delivery from clinical notes. For earliest delivery recorded in the EHR, we assigned a chart-review based label according to the gestational age at delivery (<37 weeks: preterm; 37 and 42 weeks: term, ≥42 weeks: postterm). The precision/positive predictive value for the ascertained delivery type as a binary variable ('preterm' or 'not-preterm') was calculated using the chart reviewed label as the gold standard. To compare the ascertainment strategy to a simpler phenotyping algorithm, we compared the concordance of the label derived from delivery-ICDs to one based on the gestational age within three days of delivery. This simpler phenotyping approach resulted in a lower PPV (85%) and recall (93%; Fig. S1B) compared to the billing-code-based ascertainment strategy.

Training and evaluating boosted decision trees to predict preterm birth

All models for predicting preterm birth used boosted decision trees as implemented in XGBoost v0.82 *(30)*. Unless stated otherwise, we trained models to predict the earliest delivery in a woman's EHR as preterm or not-preterm. The delivery cohort was randomly split into training (80%) and held-out (20%) sets with equal proportion of preterm cases. For all models we excluded ICD-9, CPT codes, and EGA used to ascertain delivery type and date. On the training set, we use tree of Parzen estimators as implemented in hyperopt v0.1.1 *(47)* to optimize hyperparameters by maximizing the mean average precision. The best set of hyperparameters was selected after 1,000 trials using 3-fold cross-validation over the training set (80:20 split with equal proportion of preterm cases). We evaluated the performance of all models on the held-out set using Scikit-learn v0.20.2 *(48)*. All performance metrics reported are on the held-out set. For precision-recall curves, we define baseline chance for each model as the prevalence of preterm cases. To ensure no data leaks were present in our training protocol, we trained and evaluated a model using a randomly generated dataset (n=1,000 samples) with a 22% preterm prevalence. As expected, this model did not do better than chance (AUC=0.50, PR-AUC=0.22, data not show). All trained models with their optimized hyperparameters are provided at https://github.com/abraham-abin13/ptb_predict_ml.

Identifying preterm births using features derived from EHRs

As a first step, we evaluated whether billing codes could discriminate between delivery types. Models were trained to predict preterm birth using the total counts of each ICD-9, CPT, or ICD-9 and CPT code across a woman's EHR. We excluded any codes used to ascertain delivery type or date. All three models were trained and evaluated on the same cohort of women who had at least one ICD-9 and CPT code.

In addition to billing codes, we extracted structured and unstructured features from the EHRs (Fig. 3A). Structured data included self or third-party reported race (Fig. 1E), age at delivery, past medical and family history (92 features, see Supplementary Materials and Methods), and clinical labs. For training models, we only included clinical labs obtained during the first pregnancy and excluded values greater than four standard deviations from the mean. To capture the trajectory of each clinical lab (307 clinical labs, see Supplementary Materials and Methods), we trained models using the mean, median, minimum, and maximum values across the pregnancy as features. For unstructured clinical text in obstetric and nursing clinical notes, we applied CLAMP *(49)* to extract UMLS (Unified Medical Language System) concepts unique

identifiers (CUIs and included those with positive assertions with > 0.5% frequency across all EHRs). When training preterm birth prediction models, we one-hot encoded categorical features. No transformations were applied to the continuous features.

A subset of women (n=905) was genotyped on the Illumina MEGA^{EX} platform. We applied standard GWAS quality control steps *(50)* using PLINK v1.90b4s *(51)*. We calculated a polygenic risk score for each white woman with genotype data based on the largest available preterm birth GWAS *(24)* using PRSice-2 *(52, 53)*. We assumed an additive model and summed the number of risk alleles at single nucleotide polymorphisms (SNPs) weighted by their strength of association with preterm birth (effect size). PRSice determined the optimum number of SNPs by testing the polygenic risk score for association with preterm birth in our delivery-cohort at different GWAS p-value thresholds. We included date of birth and five genetic principal components to control for ancestry. Our final polygenic risk score used 356 preterm birth associated SNPs (GWAS p-value < 0.00025).

Next, we evaluated whether adding EHR features could improve preterm birth prediction. Since the number of women varied across EHR feature, we created subsets of the delivery cohort for each EHR feature. Each subset included women with at least one recorded value for the EHR feature and billing codes. Then we trained three models as described above for each subset: 1) using only the EHR feature being evaluated, 2) using ICD-9 & CPT codes, and 3) using the EHR feature with ICD-9 & CPT codes. Thus, all three models for a given EHR feature were trained and evaluated on the same cohort of deliveries (Fig. 3A).

Predicting preterm birth before delivery using billing codes and clinical risk factors

In addition to training models using features across a woman's EHR, we also evaluated models using features present before delivery. Subtracting the estimated gestational age (recorded within three days of delivery) from the date of delivery, we obtained date of conception. Next, we trained models using ICD-9 and CPT codes timestamped before different gestational timepoints (Fig. 4A): 0, 13, 28, 32, and 35 weeks of gestation. We compared the performance of models using only billing codes to clinical risk factors obtained from the EHR. All risk factors were encoded as binary features. Risk factors such as diabetes status, fetal abnormalities, and sickle cell disease status was defined based on at least one corresponding ICD-9 code occurring before the date of delivery. The remaining risk factors, such as race (Black, Asian, or Hispanic was encoded as higher risk), age at delivery (> 34 or <18 years old), pre-pregnancy BMI ≥ 35, and pre-pregnancy hypertension (>120/80), were extracted from structured fields in EHR. Pre-pregnancy value was defined as the most recent measurement occurring before nine months of the delivery date. The association between risk factors and preterm birth was evaluated using a chi-squared test of independence implemented in SciPy v1.2.0 *(54)*.

In addition to evaluating models based on the date of conception, we trained models at different timepoints before the date of delivery (Fig. S3) using the same cohort of women by requiring every individual in this cohort had to have at least one ICD-9 or CPT code before each timepoint. Evaluating models before the date of delivery increased the sample size (n=15,481) compared to a prospective conception-based design (n=12,410) and yielded similar results.

Evaluating model performance on spontaneous preterm births, by delivery type, and recurrent preterm birth

We compared how models trained used billing codes (ICD-9 &CPT) performed in different clinical contexts. First, we evaluated the accuracy of predicting spontaneous preterm birth using models trained to predict all types of preterm births. From all preterm cases in the held-out set, we excluded women who met any of the following criteria to create a cohort of spontaneous preterm births: medically induced labor, delivery by cesarean section, or preterm premature rupture of membranes. The ICD-9 and CPT codes used to identify exclusion criteria are provided in Supplementary Materials and Methods. We calculated recall/sensitivity as the number of predicted spontaneous preterm births out of all spontaneous preterm births in the held-out set. We used the same approach to quantify performance of models trained using clinical risk factors (Fig. 4C).

We trained models to predict preterm birth among cesarean sections and vaginal deliveries separately using billing codes (ICD-9 & CPT) as features. Deliveries were labeled as cesarean sections or vaginal deliveries if they had at least one relevant billing code (ICD-9 or CPT) occurring within ten days of the date of first delivery in EHR. Billing codes used to determine delivery type are provided in Supplementary Materials and Methods. Deliveries with billing codes for both cesarean and vaginal deliveries were excluded. We trained separate models to predict cesarean and vaginal deliveries (Fig. 5 and Fig. S8).

We evaluated how well models using billing codes could predict recurrent preterm birth. From our delivery cohort, we retained women whose first delivery in the EHR was preterm and a second delivery for which we ascertained the type (preterm vs. not-preterm) as described above for the first delivery. We trained models using billing codes (ICD-9 & CPT) at timepoints before the date of delivery because the majority of this cohort did not have reliable EGA at the second delivery. As described earlier, separate models were trained using billing codes timestamped before timepoint being evaluated.

Preterm birth prediction in independent UCSF cohort

We evaluated how well models trained at Vanderbilt using billing codes would replicate in an external cohort assembled at UCSF. Only the first delivery in the EHR was used for prediction. Women with twins or multiple gestations, identified using billing codes (Supplementary Materials and Methods), were excluded. Delivery type (preterm vs. not preterm) was assigned based on the presence of ICD-10 codes. Term (or not-preterm) deliveries were determined by the presence of an ICD-10 code beginning with the characters “O80”, specifying an encounter for full-term delivery. Preterm deliveries were determined by both the absence of ICD-10 codes beginning with “O80” and the presence of codes beginning with “O60.1”, the family of codes for preterm labor with preterm delivery. We trained models using ICD-9 codes present before 28 weeks of gestation on the Vanderbilt cohort to predict preterm birth. CPT codes were not used since they were not available from the UCSF EHR system. The 28-week model was evaluated on the Vanderbilt held-out set and the independent UCSF cohort.

Feature interpretation from boosted decision tree models

To determine feature importance, we used SHapley Additive exPlanation values (SHAP) *(36, 37, 55)* to determine the marginal additive contribution of each feature. For the held-out Vanderbilt cohort and the UCSF cohort, a SHAP value was calculated for each feature per individual. Feature importance was summarized by taking the mean of the absolute value of SHAP scores across individuals. The top fifteen features based on the mean absolute SHAP value in either the Vanderbilt or UCSF cohorts values are reported. To compare how feature importance varies at Vanderbilt and UCSF, we computed the Pearson correlation of the mean absolute SHAP values.

Acknowledgments: We thank the members of the Capra lab for thoughtful discussion on this project.

Funding: AA was supported by the American Heart Association fellowship 20PRE35080073, National Institutes of Health (NIH, T32GM007347), the March of Dimes, and the Burroughs Wellcome Fund. MS, IK, and BL were supported by the March of Dimes and NIH (NLM K01LM012381). JAC was supported by the NIH(R35GM127087), March of Dimes, and the Burroughs Wellcome Fund. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the March of Dimes, or the Burroughs Wellcome Fund.

Author contributions: *Conceptualization, Methodology:* A.A. and J.A.C conceived and designed the study. *Data curation:* C.A.B. extracted billing codes, clinical notes, and performed concept extraction on the Vanderbilt cohort. P.S and L.K.D extracted, cleaned, and provided clinical laboratory data during pregnancy on the Vanderbilt cohort. *Resources:* D.R.V provided obstetric and nursing notes on the Vanderbilt cohort. B.L., I.K, MS extracted the delivery cohort from UCSF. *Formal Analysis, Investigation:* A.A. performed all analyses on the Vanderbilt cohort under supervision from J.A.C. B.L. and I.K evaluated models on UCSF cohorts under supervision from M.S. *Funding acquisition:* J.A.C. *Writing:* A.A. wrote the manuscript with guidance from J.A.C, M.S., L.M. and A.R.

Competing interests: LJM is a consultant for Mirvie, Inc.

Data and materials availability: All code and models in this study are available at https://github.com/abraham-abin13/ptb_predict_ml.

References:

1. R. L. Goldenberg, J. F. Culhane, J. D. Iams, R. Romero, Epidemiology and causes of preterm birth, *The Lancet* **371**, 75–84 (2008).

2. H. Blencowe, S. Cousens, M. Z. Oestergaard, D. Chou, A.-B. Moller, R. Narwal, A. Adler, C. V. Garcia, S. Rohde, L. Say, J. E. Lawn, National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications, *The Lancet* **379**, 2162–2172 (2012).

3. F. C. Barros, A. T. Papageorghiou, C. G. Victora, J. A. Noble, R. Pang, J. Iams, L. C. Ismail, R. L. Goldenberg, A. Lambert, M. S. Kramer, M. Carvalho, A. Conde-Agudelo, Y. A. Jaffer, E. Bertino, M. G. Gravett, D. G. Altman, E. O. Ohuma, M. Purwar, I. O. Frederick, Z. A. Bhutta, S. H. Kennedy, J. Villar, The Distribution of Clinical Phenotypes of Preterm Birth Syndrome, *JAMA Pediatrics* **169**, 220–10 (2015).

4. W. M. Callaghan, M. F. MacDorman, S. A. Rasmussen, C. Qin, E. M. Lackritz, The Contribution of Preterm Birth to Infant Mortality Rates in the United States, *Pediatrics* **118**, 1566–1573 (2006).

5. L. Liu, S. Oza, D. Hogan, Y. Chu, J. Perin, J. Zhu, J. E. Lawn, S. Cousens, C. Mathers, R. E. Black, Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals, *The Lancet* **388**, 3027–3035 (2016).

6. R. Romero, S. Dey, S. Fisher, Preterm labor: One syndrome, many causes, *Science* **345**, 760–765 (2014).

7. J. Iams, R. Goldenberg, P. Meis, B. Mercer, A. Moawad, A. Das, E. Thom, D. McNellis, R. Copper, F. Johnson, J. Roberts, The Length of the Cervix and the Risk of Spontaneous Premature Delivery, *New Engl J Medicine* **334**, 567–573 (1996).

8. F. Fuchs, B. Monet, T. Ducruet, N. Chaillet, F. Audibert, Effect of maternal age on the risk of preterm birth: A large cohort study, *Plos One* **13**, e0191002 (2018).

9. B. M. Mercer, R. L. Goldenberg, A. H. Moawad, P. J. Meis, J. D. Iams, A. F. Das, S. N. Caritis, M. Miodovnik, M. K. Menard, G. R. Thurnau, M. P. Dombrowski, J. M. Roberts, D. McNellis, F. the N. I. of C. H. H. D. M.-F. M. U. Network, The Preterm Prediction Study: Effect of gestational age and cause of preterm birth on subsequent obstetric outcome, *Am J Obstet Gynecol* **181**, 1216–1221 (1999).

10. S. Mazaki-Tovi, R. Romero, J. P. Kusanovic, O. Erez, B. L. Pineles, F. Gotsch, P. Mittal, N. G. Than, J. Espinoza, S. S. Hassan, Recurrent Preterm Birth, *Semin Perinatol* **31**, 142–158 (2007).

11. C. V. Ananth, R. S. Kirby, A. M. Vintzileos, Recurrence of preterm birth in twin pregnancies in the presence of a prior singleton preterm birth, *J Maternal-fetal Neonatal Medicine* **21**, 289–295 (2008).

12. L. J. Muglia, M. Katz, The Enigma of Spontaneous Preterm Birth, *New England Journal of Medicine* **362**, 529–535 (2010).

13. N. Auger, T. U. N. Le, A. L. Park, Z.-C. Luo, Association between maternal comorbidity and preterm birth by severity and clinical subtype: retrospective cohort study, *BMC Pregnancy and Childbirth* **11**, 75 (2011).

14. M. Carter, S. Fowler, A. Holden, E. Xenakis, D. Dudley, The Late Preterm Birth Rate and Its Association with Comorbidities in a Population-Based Study, *American Journal of Perinatology* **28**, 703–708 (2011).

15. L. Francesca, M. Laura, R. Giuseppe, D. A. Francesco, B. Ersilia, P. Leonardo, A. Domenico, Biomarkers for predicting spontaneous preterm birth: an umbrella systematic review, *The Journal of Maternal-Fetal & Neonatal Medicine* **0**, 726–734 (2019).

16. T. T. M. Ngo, M. N. Moufarrej, M.-L. H. Rasmussen, J. Camunas-Soler, W. Pan, J. Okamoto, N. F. Neff, K. Liu, R. J. Wong, K. Downes, R. Tibshirani, G. M. Shaw, L. Skotte, D. K. Stevenson, J. R. Biggio, M. A. Elovitz, M. Melbye, S. R. Quake, Noninvasive blood tests for fetal development predict gestational age and preterm delivery., *Science* **360**, 1133–1136 (2018).

17. R. J. Baer, M. R. McLemore, N. Adler, S. P. Oltman, B. D. Chambers, M. Kuppermann, M. S. Pantell, E. E. Rogers, K. K. Ryckman, M. Sirota, L. Rand, L. L. Jelliffe-Pawlowski, Pre-pregnancy or first-trimester risk scoring to identify women at high risk of preterm birth, *European Journal of Obstetrics and Gynecology* **231**, 235–240 (2018).

18. A. Weber, G. L. Darmstadt, S. Gruber, M. E. Foeller, S. L. Carmichael, D. K. Stevenson, G. M. Shaw, Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women., *Annals of epidemiology* **28**, 783-789.e1 (2018).

19. J. M. Schaaf, A. C. J. Ravelli, B. W. J. Mol, A. Abu-Hanna, Development of a prognostic model for predicting spontaneous singleton preterm birth., *European journal of obstetrics, gynecology, and reproductive biology* **164**, 150–155 (2012).

20. Y. Dabi, S. Nedellec, C. Bonneau, B. Trouchard, R. Rouzier, A. Benachi, Clinical validation of a model predicting the risk of preterm delivery., *PloS one* **12**, e0171801 (2017).

21. N. H. Morken, K. Källen, B. Jacobsson, Predicting Risk of Spontaneous Preterm Delivery in Women with a Singleton Pregnancy, *Paediatric and Perinatal Epidemiology* **28**, 11–22 (2014).

22. N. Suff, L. Story, A. Shennan, The prediction of preterm delivery: What is new?, *Semin Fetal Neonat M* **24**, 27–32 (2018).

23. N. S. Abul-Husn, E. E. Kenny, Personalized Medicine and the Power of Electronic Health Records, *Cell* **177**, 58–69 (2019).

24. G. Zhang, B. Feenstra, J. Bacelis, X. Liu, L. M. Muglia, J. Juodakis, D. E. Miller, N. Litterman, P.-P. Jiang, L. Russell, D. A. Hinds, Y. Hu, M. T. Weirauch, X. Chen, A. R. Chavan, G. P. Wagner, M. Pavličev, M. C. Nnamani, J. Maziarz, M. K. Karjalainen, M. Rämet, V. Sengpiel, F. Geller, H. A. Boyd, A. Palotie, A. Momany, B. Bedell, K. K. Ryckman, J. M. Huusko, C. R. Forney, L. C. Kottyan, M. Hallman, K. Teramo, E. A. Nohr, G. D. Smith, M. Melbye, B. Jacobsson, L. J. Muglia, Genetic Associations with Gestational Duration and

Spontaneous Preterm Birth., *The New England journal of medicine* **377**, NEJMoa1612665-1167 (2017).

25. N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, A. Connell, C. O. Hughes, A. Karthikesalingam, J. Cornebise, H. Montgomery, G. Rees, C. Laing, C. R. Baker, K. Peterson, R. Reeves, D. Hassabis, D. King, M. Suleyman, T. Back, C. Nielson, J. R. Ledsam, S. Mohamed, A clinically applicable approach to continuous prediction of future acute kidney injury, *Nature* **572**, 116–119 (2019).

26. C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, *J Am Med Inform Assn* **25**, 1419–1428 (2018).

27. J. Zhao, Q. Feng, P. Wu, R. A. Lupu, R. A. Wilke, Q. S. Wells, J. C. Denny, W.-Q. Wei, Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction, *Scientific Reports* **9**, 1–10 (2019).

28. B. A. Goldstein, A. M. Navar, M. J. Pencina, J. P. A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, *Journal of the American Medical Informatics Association* **24**, 198–208 (2017).

29. A. G. Paquette, L. Hood, N. D. Price, Y. Sadovsky, Deep phenotyping during pregnancy for predictive and preventive medicine., *Science Translational Medicine* **12**, eaay1059 (2020).

30. B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, R. Rastogi, T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, *Arxiv* , 785–794 (2016).

31. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Data Mining, Inference, and Prediction, , 261–294 (2008).

32. N. S. Artzi, S. Shilo, E. Hadar, H. Rossman, S. Barbash-Hazan, A. Ben-Haroush, R. D. Balicer, B. Feldman, A. Wiznitzer, E. Segal, Prediction of gestational diabetes based on nationwide electronic health records, *Nat Med* **26**, 71–76 (2020).

33. K. M. Corey, S. Kashyap, E. Lorenzi, S. A. Lagoo-Deenadayalan, K. Heller, K. Whalen, S. Balu, M. T. Heflin, S. R. McDonald, M. Swaminathan, M. Sendak, Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study., *Plos Med* **15**, e1002701 (2018).

34. L. Jing, A. E. U. Cerna, C. W. Good, N. M. Sauers, G. Schneider, D. N. Hartzel, J. B. Leader, H. L. Kirchner, Y. Hu, D. M. Riviello, J. V. Stough, S. Gazes, A. Haggerty, S. Raghunath, B. J. Carry, C. M. Haggerty, B. K. Fornwalt, A Machine Learning Approach to Management of Heart Failure Populations., *Jacc Hear Fail* **8**, 578–587 (2020).

35. J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, , 233–240 (2006).

36. S. M. Lundberg, S.-I. Lee, in (2017), pp. 4765--4774.

37. S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, *Nature Biomedical Engineering* **2**, 749–760 (2018).

38. A. A. Creanga, C. J. Berg, C. Syverson, K. Seed, F. C. Bruce, W. M. Callaghan, Pregnancy-Related Mortality in the United States, 2006–2010, *Obstetrics Gynecol* **125**, 5–12 (2015).

39. A. Hirshberg, S. K. Srinivas, Epidemiology of maternal morbidity and mortality, *Semin Perinatol* **41**, 332–337 (2017).

40. C. Gao, S. Osmundson, D. R. V. Edwards, G. P. Jackson, B. A. Malin, Y. Chen, Deep learning predicts extreme preterm birth from electronic health records, *J Biomed Inform* **100**, 103334 (2019).

41. M. Phelan, N. A. Bhavsar, B. A. Goldstein, Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference., *Egems Wash Dc* **5**, 22 (2017).

42. J.-M. Moutquin, Classification and heterogeneity of preterm birth, *Bjog Int J Obstetrics Gynaecol* **110**, 30–33 (2003).

43. C. Phillips, Z. Velji, C. Hanly, A. Metcalfe, Risk of recurrent spontaneous preterm birth: a systematic review and meta-analysis, *Bmj Open* **7**, e015402 (2017).

44. N. H. Shah, A. Milstein, S. C. B. PhD, Making Machine Learning Models Clinically Useful, *Jama* **322**, 1351 (2019).

45. M. A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data, *Jama Intern Med* **178**, 1544 (2018).

46. C. Weng, N. Shah, G. Hripcsak, Deep Phenotyping: Embracing Complexity and Temporality—Towards Scalability, Portability, and Interoperability, *J Biomed Inform* **105**, 103433 (2020).

47. J. Bergstra, D. Yamins, D. Cox, in (2013), pp. 115--123.

48. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, D. A. and C. Passos, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12**, 2825--2830 (2011).

49. E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, H. Xu, CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines, *J Am Med Inform Assn* **25**, 331–336 (2017).

50. A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. M. Claire, E. M. Derks, A tutorial on conducting genome-wide association studies: Quality control and statistical analysis, *International Journal of Methods in Psychiatric Research* **27**, e1608 (2018).

51. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets., *GigaScience* **4**, 7 (2015).

52. J. Euesden, C. M. Lewis, P. F. O'Reilly, PRSice: Polygenic Risk Score software., *Bioinformatics* **31**, 1466–1468 (2015).

53. S. W. Choi, P. F. O'Reilly, PRSice-2: Polygenic Risk Score software for biobank-scale data, *Gigascience* **8** (2019), doi:10.1093/gigascience/giz082.

54. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de M. Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza, SciPy 1.0: fundamental algorithms for scientific computing in Python., *Nat Methods* **17**, 261–272 (2020).

55. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, *Nat Mach Intell* **2**, 56–67 (2020).

Main Figures

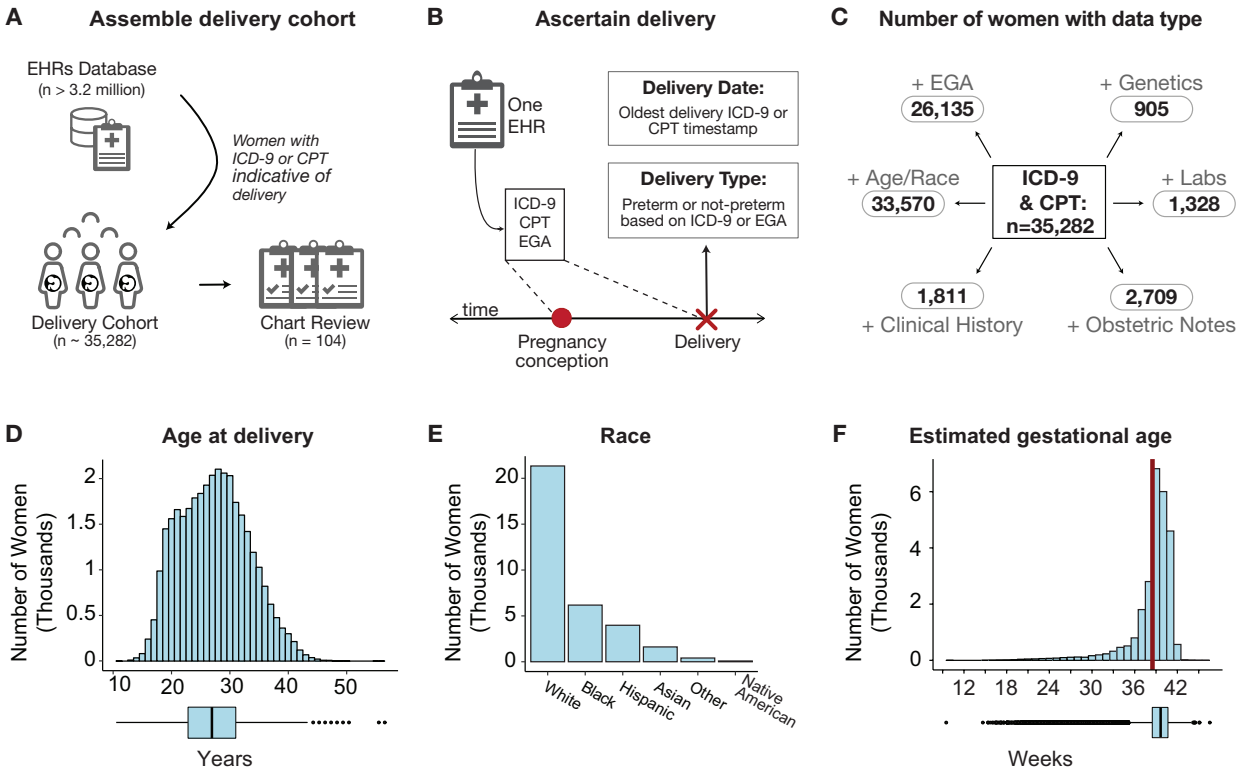

Fig. 1. **Definition and attributes of Vanderbilt delivery cohort. (A)** Schematic overview of the assembly of the delivery cohort from electronic health records (EHRs). Using billing codes, women with at least one delivery were extracted from the EHR database (n=35,282). **(B)** Delivery date and type were ascertained using ICD-9, CPT, and/or estimated gestational age (EGA) from each woman's EHR (Methods). From this cohort, 104 randomly selected EHRs were chart reviewed to validate the preterm birth label for the first recorded delivery. **(C)** Number of women in billing code cohort with estimated gestational age (+EGA), demographics (+Age, self- or third-party reported Race), clinical labs (+Labs), clinical obstetric notes (+Obstetric notes), patient clinical history (+Clinical History), and genetic data (+Genetics). **(D)** The distribution of age at first delivery in EHR (mean 27.3 years; 23.0–31.0 years, 25th and 75th percentiles). **(E)** Counts of women by self- or third-party reported race (White: 21,343; Black: 6,178; Hispanic: 3,979; Asian: 1,617; Other: 409; Native American: 84). **(F)** The EGA distribution at delivery (mean 38.5 weeks (red line); 38.0-40.3 weeks, 25th and 75th percentiles). Less than 0.015% (n=49) deliveries have EGA below 20 weeks.

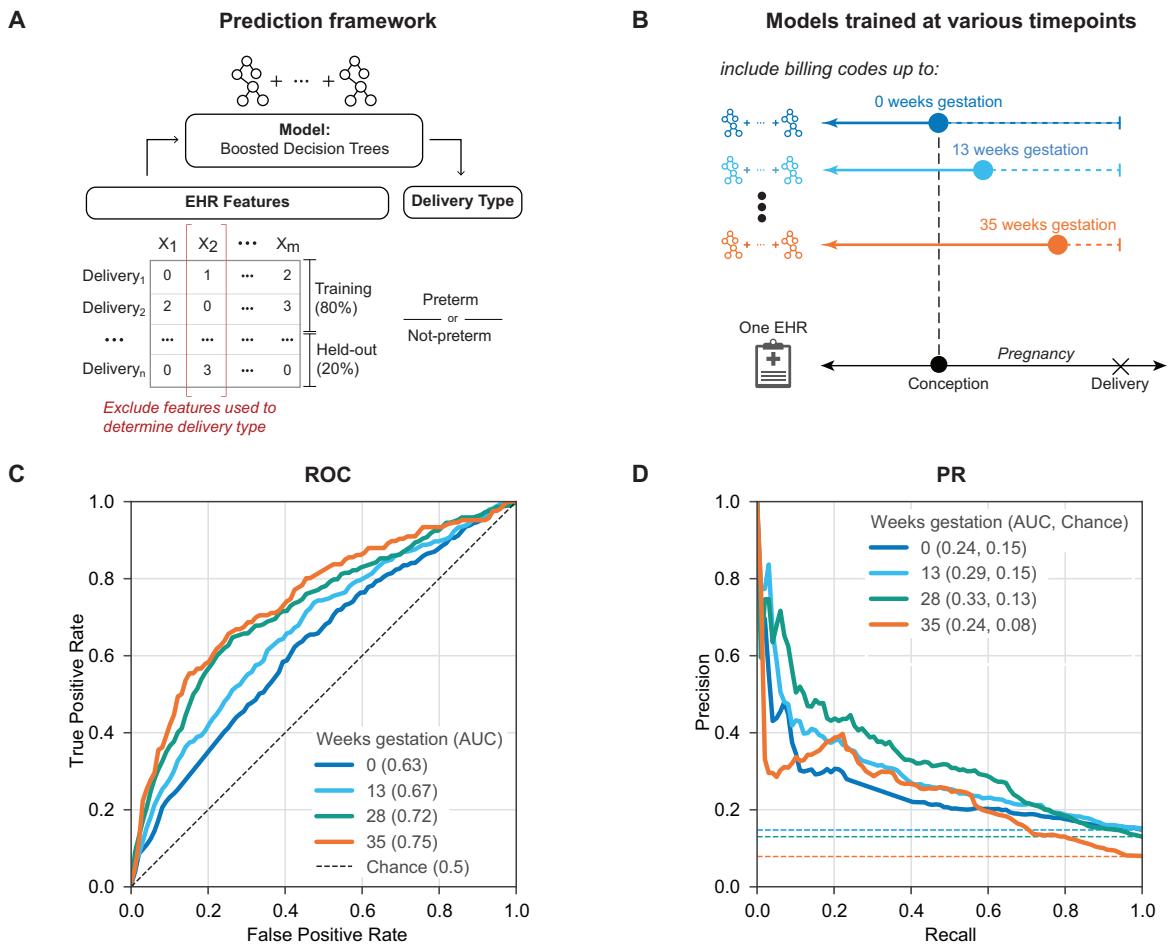

Fig. 2. Machine learning classifiers accurately identify preterm birth using billing codes present before 28 weeks of gestation. (A) Machine learning framework for training and evaluating all models. We train models (boosted decision trees) on 80% of each cohort to predict the delivery as preterm or not-preterm. EHR features used to ascertain delivery type are excluded from training. Performance is reported on 20% of the remaining held-out cohort using area under the ROC and precision-recall curves (ROC-AUC, PR-AUC). **(B)** We trained models using billing codes (ICD-9 and CPT) present before each of the following timepoints during pregnancy: 0, 13, 28, and 35 weeks of gestation. Women who already delivered were excluded at each timepoint. To facilitate comparison across timepoints, we downsampled cohorts available so that the models were trained on a cohort with similar numbers of women. **(C)** The ROC-AUC increased from conception at 0 weeks (0.63, dark blue line) to 35 weeks of gestation (0.75, orange line) compared to a chance (black dashed line) AUC of 0.5. **(D)** The model at 28 weeks of gestation achieved the highest PR-AUC (0.33). This is an underestimate of the possible performance; the accuracy improves further when all women with data available at 28 weeks are considered (Fig. 4b). Chance (dashed lines) represents the preterm birth prevalence in each cohort.

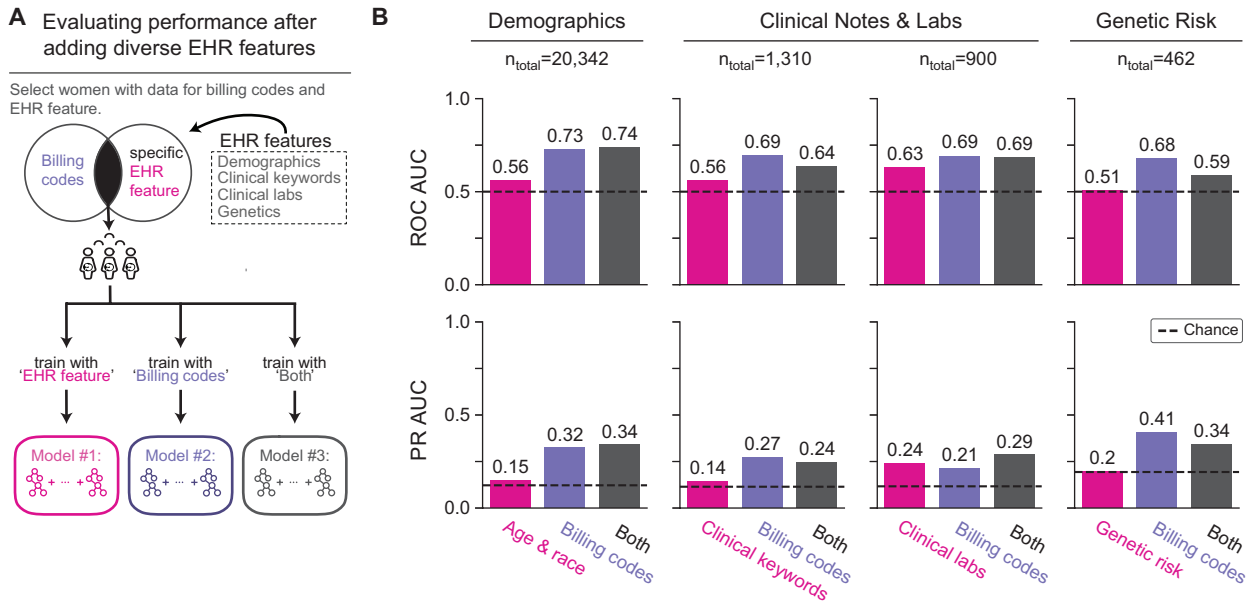

Fig. 3. Combing demographic, clinical, and genetic features does not substantially improve preterm birth prediction compared to using only billing codes. (A) Framework for evaluating change in preterm birth prediction performance after incorporating diverse types of EHR features with billing codes (ICD-9 and CPT codes). We used only features and billing codes occurring before 28 weeks of gestation. EHR features are grouped by sets of demographic factors (age and race), clinical keywords (UMLS concept unique identifiers from obstetric notes), clinical labs, and genetic risk (polygenic risk score for preterm birth). We compared three models for each feature set: 1) using only the feature set being evaluated (pink), 2) using only billing codes ('Billing codes', purple), and 3) using the feature set combined with billing codes ('Both', gray). For each feature set, we considered the subset of women who had at least one recorded value for the EHR feature and billing codes. All three models for a given EHR feature set considered the same pregnancies, but there are differences in the cohorts considered across features set due to differences in data availability; n_{total} is the number of women (training and held-out) for each feature set. **(B)** Each of the three models (x-axis) and their ROC-AUC and PR-AUC (y-axes) are shown. Each of the additional EHR features performed worse than the billing codes only model and did not substantially improve performance when combined with the billing codes. Dotted lines represent chance of 0.5 for ROC-AUC and the preterm birth prevalence for PR-AUC. Even when including EHR features before and after delivery in this framework revealed the same pattern with no substantial improvement in predictive performance compared to the billing codes only model (Fig. S6).

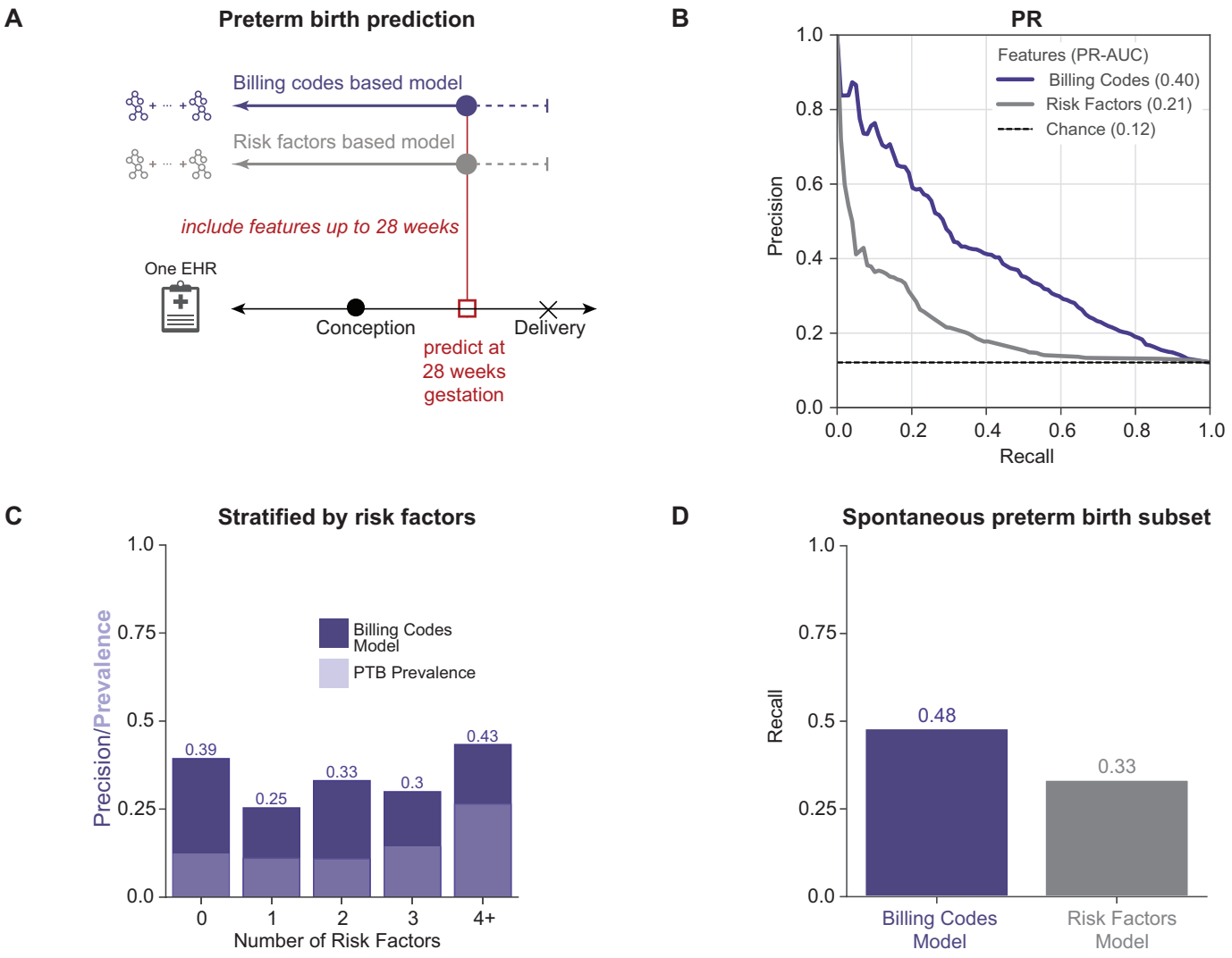

Fig. 4. Billing-code-based model outperforms a model based on clinical risk factors. (A) We compared the performance of boosted decision trees trained using either billing codes (ICD-9 and CPT) present before 28 weeks of gestation (purple) or known clinical risk factors (gray) to predict preterm delivery. Clinical risk factors (Methods) included self- or third-party reported race (Black, Asian, or Hispanic), age at delivery (> 34 or <18 years old), diabetes status, sickle cell disease status, presence of fetal abnormalities, pre-pregnancy BMI >35, and pre-pregnancy hypertension (>120/80). Both models were trained and evaluated on the same cohort of women (n=21,099). **(B)** Precision-recall curves for model using billing codes (purple line) or clinical risk factors (gray line). Preterm births are predicted more accurately by models using billing codes at 28 weeks of gestation (PR-AUC = 0.40, ROC-AUC = 0.75) than using clinical risk factors as features (PR-AUC = 0.21, ROC-AUC = 0.59; Fig. S7). Chance represents the preterm birth prevalence (dashed black line). **(C)** Billing-code-based prediction model performance stratified by number of risk factors for an individual. The billing-code-based model detects more preterm cases and has higher precision (dark purple, precision ≥ 0.25) across all numbers of risk factors compared to preterm (PTB) prevalence (light purple). **(D)** The model using billing codes also performs well at predicting the subset of spontaneous preterm births in the held-out set (recall=0.48, n=75) compared to risk factors (recall=0.33).

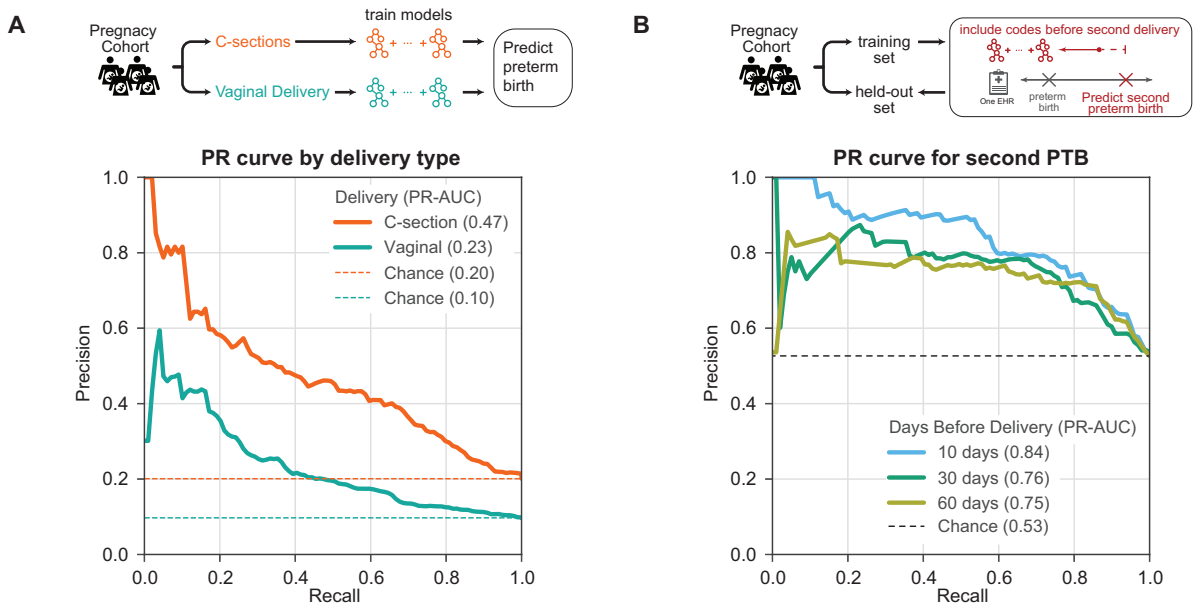

Fig. 5. Preterm birth prediction accuracy is influenced by clinical context. (A) Preterm birth prediction models trained and evaluated only on cesarean section (C-section) deliveries perform better (PR-AUC=0.47) than those trained only on vaginal delivery (PR-AUC=0.23). Billing codes (ICD-9 and CPT) present before 28 weeks of gestation were used to train a model to distinguish preterm from non-preterm birth for either C-sections (n=5,475) or vaginal deliveries (n=15,487). **(B)** Recurrent preterm birth can be accurately predicted from billing codes. We trained models to predict preterm birth for a second delivery in a cohort of 1,416 high-risk women with a prior preterm birth documented in their EHR. Three models were trained using data available at 10 days, 30 days, and 60 days before the date of second delivery. Models accurately predict the birth type in this cohort of women with a history of preterm birth (PR-AUC≥0.75). ROC-AUC varied from 0.82 at 10 days to 0.77 at 60 days before second delivery (Fig. S9). Expected performance by chance is the preterm birth prevalence in each cohort (dashed lines).

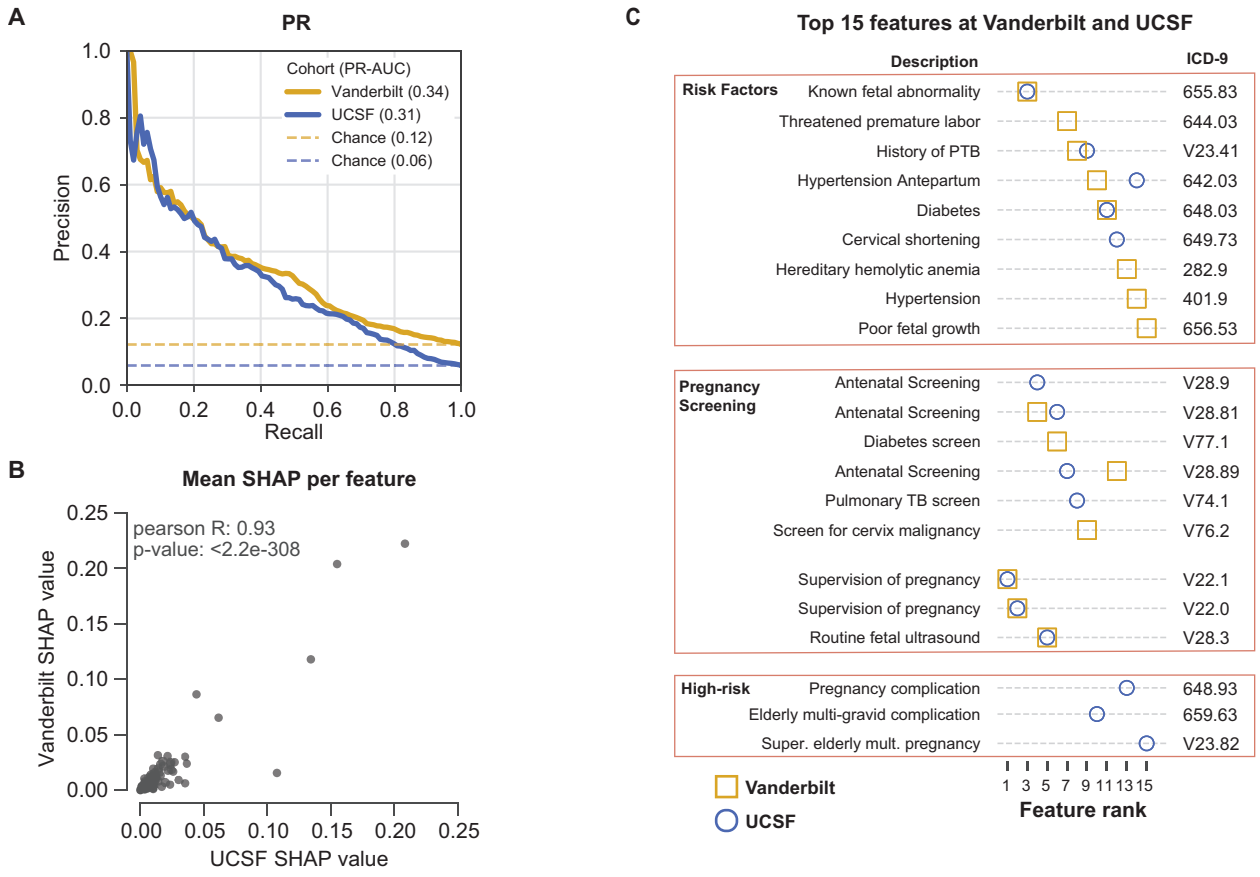

Fig. 6. Preterm birth prediction models accurately generalize to an independent cohort. **(A)** Performance of preterm birth prediction models trained at Vanderbilt applied to UCSF cohort. Models were trained using ICD-9 codes present before 28 weeks of gestation at Vanderbilt on 16,857 of women and evaluated on a held out set at Vanderbilt (n=4,215, gold) and UCSF cohort (n=5,978, blue). Models perform well at Vanderbilt (PR-AUC=0.34) and at UCSF (PR-AUC=0.31) compared to baseline prevalence (dotted lines, 0.12 and 0.06 at Vanderbilt and UCSF respectively). Note that in contrast to models presented previously this one was trained only on ICD-9 codes, due to the lack of CPT codes in the UCSF cohort. **(B)** Feature importance was estimated by the mean absolute SHapley Additive exPlanation (SHAP) value per feature in each cohort (x and y-axes). The feature importance estimates have a high positive correlation between cohorts (Pearson r=0.93 p<2.2e-308). **(C)** The top 15 features with the highest mean absolute SHAP score in the Vanderbilt cohort (gold square) or UCSF cohort (blue circle). The majority of the features were shared across cohorts and capture known risk factors (fetal abnormalities, history of preterm birth, etc.), pregnancy screening visits, and supervision of high-risk pregnancies.

Supplementary Materials:

Materials and Methods

Delivery-specific ICD-9/10 codes used to ascertain delivery type.
The following ICD-9/10 codes were used to ascertain delivery type as described in the Methods section.

- Preterm Birth: 'O60.1 ', 'O60.10', 'O60.10X0', 'O60.10X1', 'O60.10X2', 'O60.10X3', 'O60.10X4', 'O60.10X5', 'O60.10X9', 'O60.12', 'O60.12X0', 'O60.12X1', 'O60.12X2', 'O60.12X3', 'O60.12X4', 'O60.12X5', 'O60.12X9', 'O60.13', 'O60.13X0', 'O60.13X1', 'O60.13X2', 'O60.13X3', 'O60.13X4', 'O60.13X5', 'O60.13X9', 'O60.14', 'O60.14X0', 'O60.14X1', 'O60.14X2', 'O60.14X3', 'O60.14X4', 'O60.14X5', 'O60.14X9', '644.2', '644.20', '644.21'
- Term Birth: 'O60.20', 'O60.20X0', 'O60.20X1', 'O60.20X2', 'O60.20X3', 'O60.20X4', 'O60.20X5', 'O60.20X9', 'O60.22', 'O60.22X0', 'O60.22X1', 'O60.22X2', 'O60.22X3', 'O60.22X4', 'O60.22X5', 'O60.22X9', 'O60.23', 'O60.23X0', 'O60.23X1', 'O60.23X2', 'O60.23X3', 'O60.23X4', 'O60.23X5', 'O60.23X9', 'O80', 'O48.0', '650', '645.1', '645.10', '645.11', '645.13', '649.8', '649.81', '649.82'
- Postterm Birth: 'O48.1', '645.2', '645.20', '645.21', '645.23', '645.00', '645.01', '645.03'

Delivery-specific CPT codes used to ascertain delivery type.
The following CPT codes were used to ascertain delivery date: '59400', '59409', '59410', '59414', '59510', '59514', '59515', '59525', '59610', '59612', '59614', '59618', '59620', '59622'.

Identifying multiple gestations using billing codes.
Pregnancies with multiple gestations were identified using the presence of any of the following billing codes. For singleton only analyses, we excluded women with multiple gestation.

- ICD-9 Multiple Gestations: '651','651.7','651.70','651.71','651.8','651.81','651.83','651.9','651.91','651.93','652.6','652.60','652.61','652.63','V91','V91.9','V91.90','V91.91','V91.92','V91.99','651','651.0','651.00','651.01','651.03','651.1','651.10','651.11','651.13','651.2','651.20','651.21','651.23','651.3','651.30','651.31','651.33','651.4','651.40','651.41','651.43','651.5','651.50', '651.51', '651.53', 'V91','V91.0','V91.00','V91.01','V91.02','V91.03','V91.09','V91.1','V91.10','V91.11','V91.12','V91.19','V91.2','V91.20','V91.21','V91.22','V91.29','V91.9','V91.90','V91.91','V91.92', 'V91.99'
- CPT Twin codes: '74713','76802','76810','76812','76814'
- ICD-10 Multiple Gestations: 'BY4BZZZ','BY4DZZZ','BY4GZZZ','O30.801','O30.802','O30.803','O30.809','O30.811','O30.812','O30.813','O30.819','O30.821','O30.822','O30.823','O30.829','O30.891','O30.892','O30.893','O30.899','O30.91','O30.92','O30.93','O31.BX10','O31.BX11','O31.BX12','O31.BX13','O31.BX14','O31.BX15','O31.BX19','O31.BX20','O31.BX21','O31.BX22','O31.BX23','O31.BX24','O31.BX25','O31.BX29','O31.BX30','O31.BX31','O31.BX32','O31.B

X33','O31.BX34','O31.BX35','O31.BX39','O31.BX90','O31.BX91','O31.BX92','O31.BX93','O31.BX94','O31.BX95','O31.BX99'

Past medical and family history extracted from EHRs used to predict preterm birth.
The following past medical and family history features were extracted from EHRs for women with at least one recorded delivery at Vanderbilt Hospital.

- Maternal History:

'Abortion', 'Alcohol ', 'Baby's father had a child with birth defect not listed', 'Baby's father's family has history of birth defect not listed', 'Drugs ', 'Endocrine Metabolic Patient ', 'Endocrine metaboloic Patient History ', 'Gravidity', 'Hematoligic ', 'Maternal metabolic or endocrine disorders (Diabetes, PKU) ', 'Menses every 28 to 30 days ', 'Patient History Breast Disease ', 'Patient History Congential Heart Defect ', 'Patient History Cystic Fibrosis ', 'Patient History Down Syndrome ', 'Patient History GI Problems ', 'Patient History Genetic other', 'Patient History Gyn Problems ', 'Patient History Heart Disease ', 'Patient History Hemophilia or other blood disorders ', 'Patient History Huntington's Chorea ', 'Patient History Hypertension ', 'Patient History Immune or Infectious Disease ', 'Patient History Infertility or Recurrent Spontaneous Abortions ', 'Patient History Malignancies ', 'Patient History Mental Retardation ', 'Patient History Multiple births ', 'Patient History Muscular Dystrophy ', 'Patient History Neural Tube Defect ', 'Patient History Neurological Disorder ', 'Patient History Operations or Accidents ', 'Patient History Other Hospitalizations ', 'Patient History Other ', 'Patient History Other inherited or chromosomal disorders ', 'Patient History Other structural Birth defect ', 'Patient History Phlebitis or varicocities ', 'Patient History Pulmonary Disease ', 'Patient History Recurrent Pregnancy loss defined as more than 2 or stillbirth', 'Patient History STDs ', 'Patient History Sickle Cell Disease (African or Carribean American) ', 'Patient History Thalessemia (Italian, Greek, Mediterranean, or Asian Background); MCV <80 ', 'Patient History Tobacco, Alcohol, Drugs ', 'Patient History Urinary tract problems including UTIs and Pyel ', 'Patient History of Seizure', 'Patient History of sexual/physical abuse or trauma ', 'Patient's age greater than 34 at delivery ', 'Pregnancy Induced Hypertension', 'Prior Preterm_births', 'Regular exercise ', 'Term_births', 'Tobacco ', 'Urinary tract infection', 'Live_Children'

- Family History:

'Familly History Thalessemia (Italian, Greek, Mediterranean, or Asian Background); MCV <80 ', 'Family History Breast Disease ', 'Family History Congential Heart Defect ', 'Family History Cystic Fibrosis ', 'Family History Down Syndrome ', 'Family History GI Problems ', 'Family History Genetic other', 'Family History Gyn Problems ', 'Family History Heart Disease ', 'Family History Hemophilia or other blood disorders ', 'Family History Huntington's Chorea ', 'Family History Hypertension ', 'Family History Immune or Infectious Disease ', 'Family History Infertility or Recurrent Spontaneous Abortions ', 'Family History Jewish, Cajun, French Canadian (Tay Sachs) ', 'Family History Jewish: Canavan Disease, Gauchers ', 'Family History Malignancies ', 'Family History Mental Retardation ', 'Family History Metabolic or endocrine disorders (Diabetes, PKU) ', 'Family History Multiple births ', 'Family History Muscular Dystrophy ', 'Family History Neural Tube Defect ', 'Family History Neuroligcal Disorder ', 'Family History Operations or Accidents ', 'Family History Other Hospitalizations ', 'Family History Other ', 'Family History Other inherited or chromosomal disorders ', 'Family History

Other structural Birth defect ', 'Family History Phlebitis or varicocities ', 'Family History Pulmonary Disease ', 'Family History Recurrent Pregnancy loss defined as more than 2 or stillbirth', 'Family History STDs ', 'Family History Sickle Cell Disease (African or Carribean American) ', 'Family History Tobacco, Alcohol, Drugs ', 'Family History Urinary tract problems including UTIs and Pyel ', 'Family History of Seizure', 'Family History of sexual/physical abuse or trauma ', 'Jewish, Cajun, French Canadian (Tay Sachs) ', 'Jewish: Canavan Disease, Gauchers'

Clinical labs measured during pregnancy used to predict preterm birth

'albumin urine, lactic acid venous, cd3 #/cumm, total protein urine, glucose blood, wbc blood, eo automated abs, atyp lymphs (abs), reaction time, lmw heparin assay, rdwsd, glucose spinal fluid, control ptt, rbc folate, calcium blood, gentamicin level, urea nitrogen ur spot, mch, aldosterone, magnesium blood, mchc, factor viii activity, sodium blood, igg quantitative blood, bicarbonate (calc), hcg beta (3rd irp), dhea sulfate, hdl cholesterol, protein csf, f t4, alt blood, neutrophil %, k-time, metamyelo %, estriol unconjugated, sodium urine spot, cellano antigen, icterus index, nucleated rbc, protein total blood, eosoinophil (abs), erythropoietin, neutrophils %, immature retic fraction, zinc serum, c-peptide, imm granulocytes %, lipemia index, monocytes %, ssb (la)(ena) ab, igg, beta-hcg serum, protein urine, bedside glucose, troponin t, intact-pth, sm (smith) autoabs eia, ferritin, absolute cd8, sex hormone bind globulin, eosinophils %, protein c activity, cd8(cd3+)/cd45 #/cumm, glucose tol 50g, basophils %, wbc, albumin, mcv, gamma globulin, testosterone free, fio2, lymph %, pan t cd3 %, troponin-i, mono (abs), rheumatoid factor, quant d-dimer for dic, pcv blood, hgb a1c glycated poc, 25-hydroxy d3, eosoinophil (abs), carboxyhemoglobin, urea nitrogen blood, hgb a1c glycated, cholesterol blood, lamotrigine, cystatin-c, carbon dioxide blood, apa-igg, neutrophils %, myelocytes %, hdl cholesterol, vit e(alpha-tocopherol), glucose whole blood, calcium ionized, gamma glut trans blood, follicle stimulating hrm, total hemoglbin, creatinine g/24 hour, atyp lymphs %, wbc urine micro, nt automated abs, chloride blood, imm platelet fraction, fasting glucose, po2/fio2, sodium whole blood, ast blood, albumin/creatinine ratio, angle(alpha), rbc, vit d 1,25-dihydroxy, c3 quantitative blood, lymphs (abs), ldl cholesterol, triglycerides blood, testosterone, ed troponin-i wbld, o2 saturation, creatinine urine per day, triiodothyronine free, eosinophil %, rbc, rbc urine micro, thyroid stim hormone, anti-myeloperoxidase, c-reactive protein, deamidated gliadin iga abs, hyaline cast, ammonia, igg beta 2 glycoprotein i, progesterone rapid, vitamin d 25-oh total, t helper cd4 #/cumm, patient (pt), schedule q hr, keppra (levetiracetam), creatine kinase total, maternal alphafeto pr0, creatinine urine "spot", ret ct, creatinine urine "timed", specific gravity ua, iron blood, kappa light chain quant, lithium blood, 2 hour glucose, vancomycin level, anion gap, luteinizing hormone, iga quantitative blood, phenytoin (dilantin), methemoglobin, alpha-1 globulin, thyroglobulin serum, renin activity, c4 quantitative blood, rdw, urobilinogen, maternal weight, venous ph, % cd3, protein urine spot, carbamazepine (tegretol), hep b surface ab value, anti-protease 3, hemoglobin s, sed rate, amylase blood, ssa (ro)(ena) ab, igg, 25-hydroxy d total, total gamma globulin, adrenocorticotropic horm, retic hgb equiv, neut (abs), insulin, albumin, lymphs %, antithrombin iii act, myelocytes (abs), lymps %, nucleated rbc, alkaline phosphatase bld, # wbc\'s counted, fibrinogen, ed creatinine wbld, ph arterial, metamyelocytes (abs), kappa/lambda ratio, ret abs, beta globulin, basophils %, albumin blood, ed inr wbld, anti thyroid peroxab, tc:hdl ratio, afp tumor, vitamin a (retinol), albumin/creat ratio, patient location, ck-mb ratio, total volume, total t4, creatinine blood, absolute cd3, collection time, current gest age, apa-igm, ck blood, hemoglobin blood, max amplitude, transferrin blood, cd8(cd3+)/cd45 %, cd4:cd8

ratio, monocytes %, protein urine timed, beta globulin, dose, % cd8, estradiol, nucleated rbc#, cortisol, prolactin, lymphs (abs), granular cast, protein-s-activity, pcv blood, mono (abs), brain natriuretic peptide, fk-506 (tacrolimus), bilirubin conjugated, bilirubin total blood, chloride whole blood, 25-hydroxy d2, hemoglobin a, haptoglobin blood, folate serum, ck-mb, glucose urine, nucleated cell, absolute cd4, baso (abs), creatinine urine, scl-70 autoabs eia, infusion start time, squamous epithelial, g parameter, osmolality blood, baso (abs), vitamin b-12, hours of collection, inr, lipase blood, hemoglobin a2, potassium urine spot, factor v leiden coag, phosphorus inorganic, percent saturation, valproate(depakane), ldh blood, anti-dna(sle)current, lambda light chain quant, bcrabl/bcr ratio, free phenytoin, t helper cd4 %, % cd4, mean platelet volume, creatinine urine, ketone urine, igm quantitative blood, patient ptt, glucose body fluid, vit e(gamma-tocopherol), igm beta 2 glycoprotein i, maternal b-hcg, drvvt, protein total blood, ph urine, lymph abs, alpha-2 globulin, retinyl palminate, d-dimer (patient), lymphs %, o2 saturation (calc), uric acid blood, ldl cholesterol calc, non-hdl, triiodothyronine, total, testerone free female, potassium whole blood, deamidated gliadin igg abs, urobilinogen, monocytes %, protein /24 hour, neut (abs), tibc blood, apa-iga, platelet count, albumin urine, o2 saturation(venous, prealbumin blood, basophils %, eosinophil %

Identifying cesarean section and vaginal deliveries.

The following ICD-9 and CPT codes were used to label deliveries as a cesarean section vs. vaginal deliveries. We excluded deliveries if they had codes for both types of deliveries.

- Cesarean section: '669.7', '669.70', '669.71', '763.4', '74.0', '74.1', '74.2', '74.4', '74.9','74.99', '59510', '59514', '59515', '59618', '59620', '59622'
- Vaginal Deliveries: '59409', '59410', '59610', '59612', '59614'

Identifying spontaneous preterm births from electronic health records

From all preterm cases, we excluded women meeting any of the following criteria: medically induced labor, delivery by cesarean section, or preterm premature rupture of membranes. The following ICD-9 and CPT codes were used to identify these exclusion criteria.

- Medically induced labor: '73.01', '73.1', '73.4', '73.0', '73.09'
- Cesarean Section delivery: '669.7', '669.70', '669.71', '763.4', '74.0', '74.1', '74.2', '74.4', '74.9','74.99', '59510', '59514', '59515', '59618', '59620', '59622'
- Preterm premature rupture of membranes: '658.13','658.10','658.11'

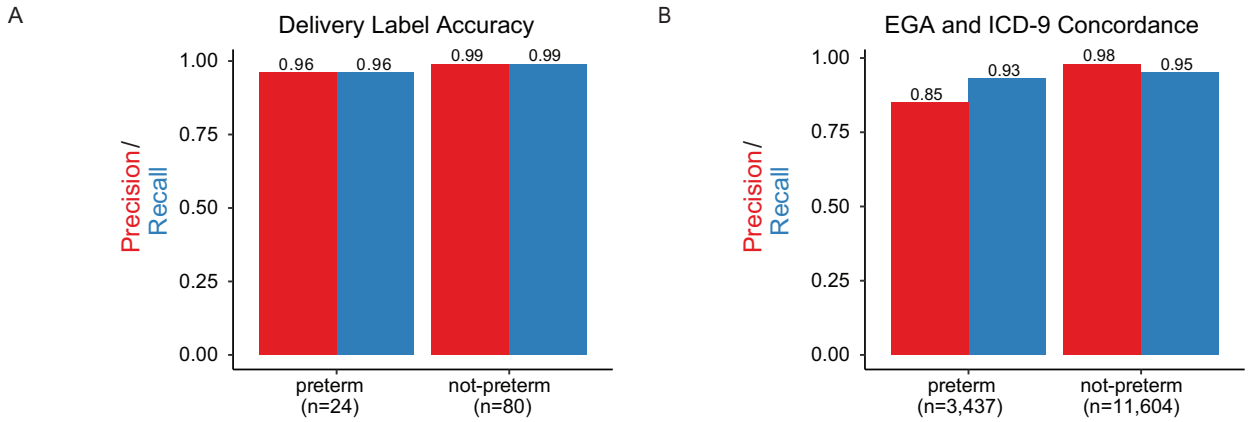

Fig. S1. Preterm births can be accurately ascertained from EHRs. (A) Accuracy of delivery type phenotyping. The phenotyping algorithm was evaluated by chart review of 104 randomly selected women. The approach has high precision and recall for a binary classification of 'preterm' or 'not-preterm'. **(B)** The concordance between estimated gestational age (EGA) within three days of delivery and ICD-9 based delivery type labels for the 15,041 women with sufficient data for both. Precision and recall values were > 93% across labels except with preterm precision (85%)

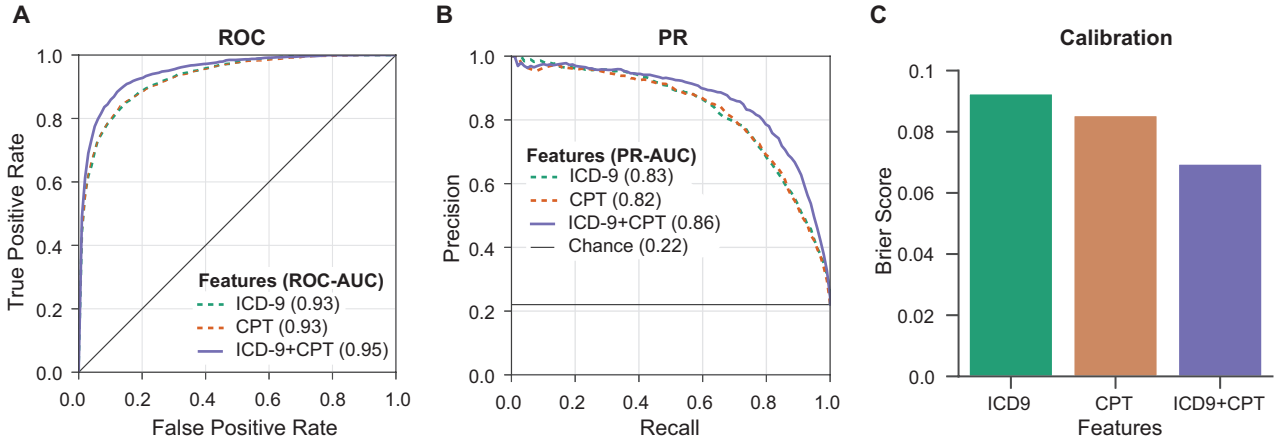

Fig. S2: Boosted decision trees trained on EHR billing codes accurately identify preterm births. We trained and validated boosted decision trees on 80% of labeled pregnancies (preterm vs. non-preterm) from the EHR cohort (n=35,282, Fig. 1). We included both singletons and multple gestations. We evaluated model performance on the held-out set using area under the ROC and precision-recall curves (ROC-AUC, PR-AUC) and Brier scores. EHR features used to ascertain delivery labels are excluded in training and evaluation of the models. **(A,B)** The boosted decision trees accurately classified deliveries by preterm birth status using only ICD-9 (green dashed line), only CPT (orange dashed line), and combined ICD-9 and CPT (solid purple) features present in a women's EHR, (ROC-AUC≥0.93, PR-AUC≥0.86). Combining ICD-9 and CPT codes achieved the best performance. **(C)** The low Brier scores (≤0.092) indicate that the models are well calibrated.

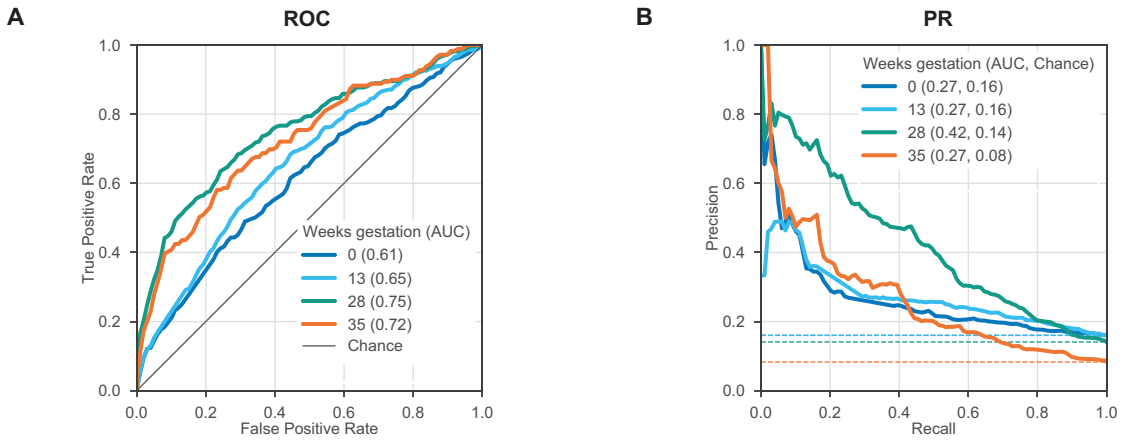

Fig. S3: Machine learning can accurately identify preterm birth including singletons and multiple gestations. We trained models (boosted decision trees) on 80% of the corresponding cohort to predict the earliest delivery as preterm or not-preterm (Methods). We included singleton and multiple gestations. Billing codes (ICD-9 and CPT) present before pregnancy (0, 13, 28, and 35 week of gestation) were used to train models. The same cohort of women (training + held-out) was used to train and evaluate across models but the sample size varied slightly (n= 11,843 to 10,799) since women who already delivered were excluded at each timepoint. **(A)** The ROC-AUC increased from conception at 0 weeks (0.61, dark blue line) to 35 weeks of gestation (0.72, orange line) compared to a chance (black line). **(B)** The model at 28 weeks of gestation achieved the highest PR-AUC (0.42). Chance (dashed lines) represents the preterm birth prevalence in each cohort.

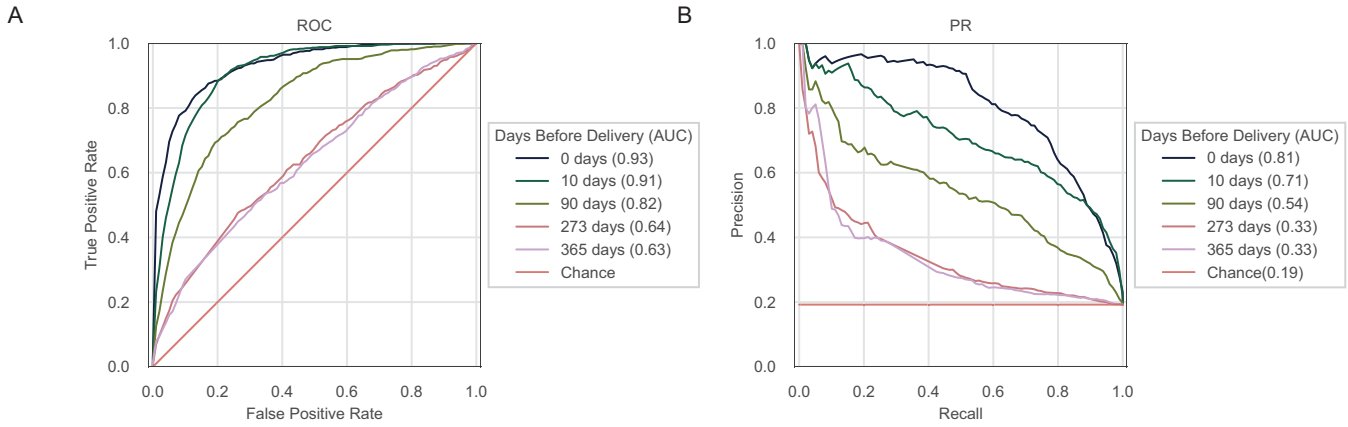

Fig. S4: Preterm birth prediction increases at timepoints closer to the date of delivery at timepoints based on days before delivery. (A) ROC and **(B)** PR curves for preterm birth prediction using billing codes (ICD-9 & CPT) at different timepoints defined from the date of delivery in the Vanderbilt cohort. Both singletons and multiple gestations are included. Chance for PR-AUC represent random prediction equal to the population prevalence of preterm birth. Model performance continuously improves as the prediction is made closer to delivery. All models are trained and evaluated on the same cohort of women (n=15,481) and the performances reported is on the held-out set (20% of cohort).

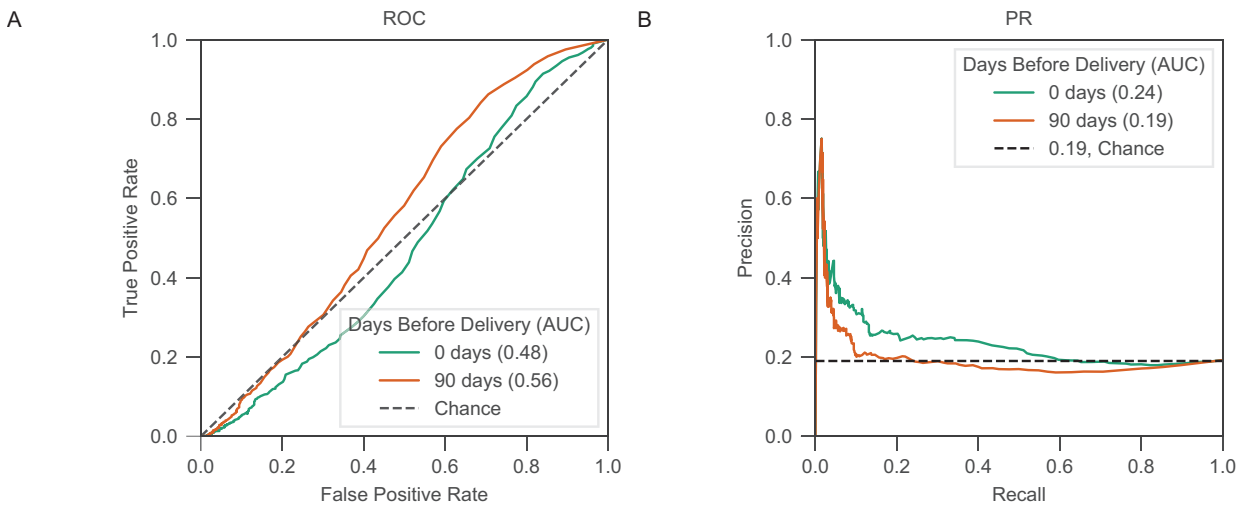

Fig. S5: Preterm birth prediction is not driven by total number of billing codes. To evaluate whether the amount of contact with the healthcare system was driving the performance of our machine learning classifiers, we assessed the discriminatory ability of the total number of billing codes (ICD-9 or CPT) in a woman's EHR to predict preterm birth. We include both singletons and multiple gestations. A simple classifier that used only the number of total billing codes preset at 0 days (green) and 90 days (orange) before the first delivery in her EHR, did not predict preterm birth better than chance **(A)** ROC-AUC = 0.562 and **(B)** PR-AUC = 0.19. The cohort consisted of the held-out set at the specified timepoints with 3,096 women.

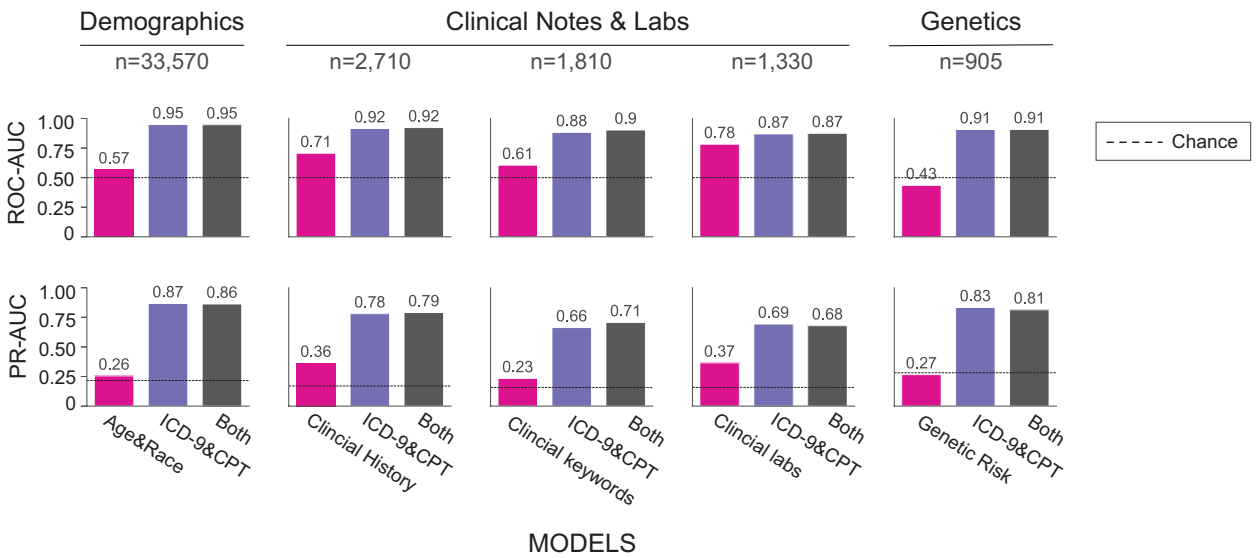

Fig. S6: Combining EHR features with billing codes do not improve model performance. We evaluated how combining EHR feature with billing codes could improve model performance. To maximize potential gains, we included billing codes and EHR features before and after delivery and included singletons and multiple gestations. EHR features are grouped in to sets of: demographic factors (age and race), clinical history (patient and familial comorbidities), clinical keywords (UMLS concept unique identifiers from obstetric notes), clinical labs, and genetic risk (polygenic risk score for preterm birth). We compared three models for each feature set: 1) using only the feature set being evaluated (pink), 2) using ICD-9 & CPT codes (purple), and 3) using the feature set combined with ICD-9&CPT codes (gray). For each feature set, we considered the subset of women who had at least one recorded value for the EHR feature and ICD&CPT codes. All three models for a given EHR feature set considered the same pregnancies, but there are differences in the cohorts considered across features set due to differences in data availability. Each of the three models (x-axis) and their ROC-AUC and PR-AUC (y-axes) are shown. Each of the additional EHR features performed worse than the billing code only model and did not substantially improve performance when combined with the billing codes. Of the other EHR features tested, clinical labs had the best predictive performance with PR-AUC of 0.37 and ROC-AUC of 0.78. Dotted lines represent chance of 0.5 for ROC-AUC and the preterm birth prevalence for PR-AUC. The total number of women (n) in each subset including the training and held out set is given.

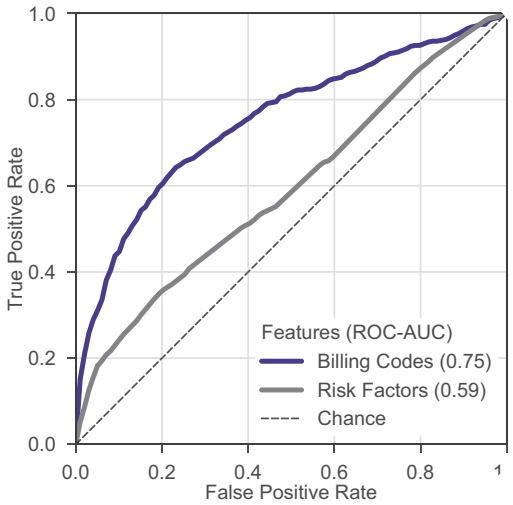

Fig. S7: Models trained using billing codes occurring before 28 weeks of gestation accurately predict preterm birth. Billing codes (ICD-9 and CPT, purple line) and clinical risk factors (dashed grey line, Methods) occurring before 28 weeks of gestation was used to train separate models to predict preterm birth. Billing-code-based model (ROC-AUC=0.75) accurately predicted preterm birth and outperformed the risk factor based model (ROC-AUC=0.59). Both models were trained and evaluated on the same cohort of women (n=21,099).

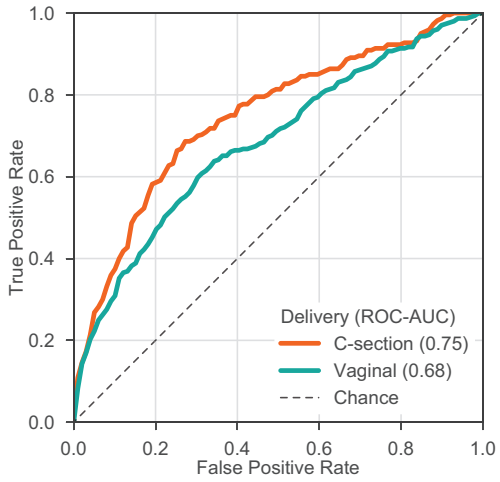

Fig. S8: Preterm birth prediction accuracy is higher for cesarean-sections compared to vaginal deliveries. After stratifying the delivery cohort into cesarean-sections (n=5,475) and vaginal (n=15,487) deliveries, we trained a model on each delivery type to predict preterm or not-preterm births. Multiple gestations were excluded. We trained models using billing codes (ICD-9 and CPT) present before 28 weeks of gestation. ROC-AUC was higher for cesarean-sections (0.75) compared to vaginal deliveries (0.68).

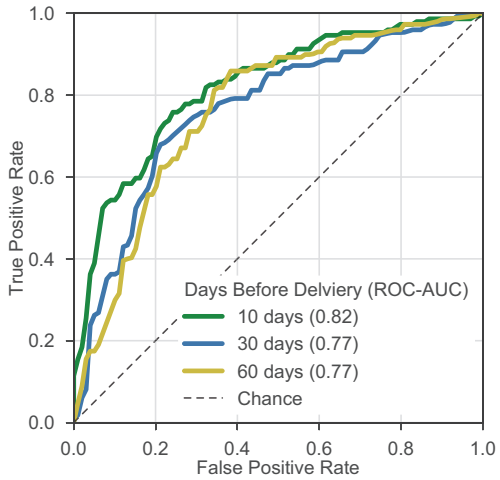

Fig. S9: Models trained using billing codes can accurately predict a second preterm birth. For women with a history of preterm birth (n=1,416, Methods), we trained models using billing codes (ICD-9 and CPT) to predict a second preterm birth. Multiple gestations were excluded. For each model, only billing codes timestamped before the specified number of days before delivery are included. Models predicted a second preterm birth accurately with the highest and lowest ROC-AUC of 0.82 at 10 days and 0.77 at 60 days before delivery respectively.

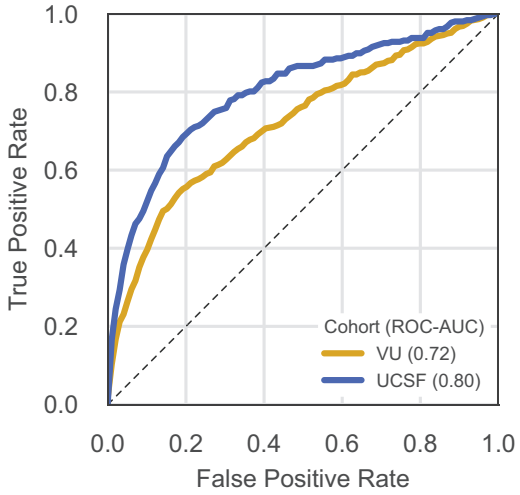

Fig. S10: Models accurately predict preterm birth and generalize to an independent cohort. Models trained using ICD-9 codes present before 28 weeks of gestation on a Vanderbilt cohort was evaluated on the held-out set at Vanderbilt (n=4,215, gold) and UCSF cohort (n=5,978, blue). Models accurately predicted preterm birth at Vanderbilt (ROC-AUC=0.72) and at UCSF (ROC-AUC=0.80).

	UCSF			Vanderbilt		
	Not-Preterm	**Preterm**	**p-value**	**Not-Preterm**	**Preterm**	**p-value**
n	5615	363		18,498	2,651	
Patient Age (mean (SD))	36.65 (5.08)	36.54 (5.96)	0.691	27.71 (5.75)	27.73 (6.38)	0.876
Patient Race (%)			<0.001			<0.001
American Indian or Alaska Native	26 (0.5)	3 (0.8)		47 (0.2)	4 (0.01)	
Asian	1,336 (23.8)	51 (14.0)		1,051 (5.8)	100 (3.8)	
Black or African American	336 (6.0)	31 (8.5)		2,962 (16.5)	486 (18.8)	
Declined	72 (1.3)	5 (1.4)		NA	NA	
Native Hawaiian/Pacific Islander	86 (1.5)	3 (0.8)		NA	NA	
Other	866 (15.4)	77 (21.2)		162 (0.9)	12 (0.04)	
Unknown	200 (3.6)	32 (8.8)		619 (3.3)	69 (2.6)	
White or Caucasian	2,693 (48.0)	161 (44.4)		11,278(63.0)	1,658 (64.2)	

Table S1: Demographic distribution of UCSF and Vanderbilt Cohort. We identified women with preterm and not-preterm deliveries at UCSF and Vanderiblt using similar ascertainment (Methods). For each women, we predicted the earliest delivery in their EHR. We report age at delivery (Patient Age) and self- or third-party reported race for both cohorts. T-tests and Chi-squared test of idenpendence was used to compare distrbutions stratfied by delivery label.