Supplemental Methods and Results:

Improving postpartum hemorrhage risk prediction using longitudinal electronic medical records

Amanda B Zheutlin\textsuperscript{1*}, Ph.D., Luciana Vieira\textsuperscript{2*}, M.D., Ryan A Shewcraft\textsuperscript{1*}, Ph.D., Shilong Li\textsuperscript{1}, Ph.D., Zichen Wang\textsuperscript{1}, Ph.D., Emilio Schadt\textsuperscript{1}, Susan Gross\textsuperscript{1,3}, M.D., Siobhan M. Dolan\textsuperscript{2,3}, M.D., M.P.H., Joanne Stone\textsuperscript{2}, M.D., Eric Schadt\textsuperscript{1,3*}, Ph.D., Li Li\textsuperscript{1,3*}, M.D., M.S.

\textsuperscript{1}Clinical Informatics, Sema4, 333 Ludlow St., Stamford, CT 06902, USA

\textsuperscript{2}Division of Maternal Fetal Medicine, Department of Obstetrics, Gynecology, and Reproductive Science, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

\textsuperscript{3}Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Supplemental Methods

Delivery cohort

Clinical data extraction, cleaning, and normalization

Digital phenotype algorithm for postpartum hemorrhage

Feature engineering

Feature selection

Model interpretation and simplification

Risk factors used in other risk assessment tools

Supplemental Results

Table S1. Training performance for machine learning risk assessment tools

Table S2. Top clinical features include novel and known risk factors

Table S3. Linear fixed effects models for case differences in vital signs prior to delivery
Table S4. Risk categories across risk assessment tools
Table S5. Test performance across risk assessment tools for EBL phenotype
Figure S1. Feature selection
Figure S2. Data availability
Figure S3. Model comparison
Supplemental Methods

Delivery cohort

We used deidentified EMR data from the MSHS, one of the largest and most comprehensive EMR systems in New York City. From the 9 million unique patients with any demographic or clinical data represented in this EMR, we derived a delivery cohort of 71,944 deliveries from 57,151 mothers occurring between January 1, 2011 and December 31, 2019. We identified deliveries using structured delivery summaries completed by the Labor & Delivery staff in the MSHS, procedure records, and encounter visit details. For all included deliveries, we extracted gestational weeks at delivery, delivery time, delivery method, and hospital admission time.

Clinical data extraction, cleaning, and normalization

Demographics, lab results, vital signs, diagnoses, medications, and procedures occurring one year prior to pregnancy (pregnancy initiation was estimated based on gestational age at delivery) and up to one year after delivery were extracted for deliveries in our cohort. Furthermore, since patients in our dataset can have multiple pregnancies, we limit the data used by our model to 8 months prior to pregnancy up to and including delivery. This limitation eliminated concern for non-independence of journeys due to patients with previous pregnancies. The shortest time between pregnancy journeys for the same patient was 295 days, roughly 9.5 months. By definition, PPH occurs after delivery, so we used delivery time as a natural cutoff for data inclusion and removed all data occurring at delivery time or later for all patients. Considering the high value of clinical information immediately preceding delivery, including during labor (intrapartum), we split available medication, vital and lab data into two time periods: pre-hospital admission and post-admission (but still prior to delivery time). Data were standardized by mapping original names and values to common Unified Medical Language
System (UMLS) frameworks to increase interoperability between healthcare systems and reduce dimensionality. All observations were cleaned and normalized within data type.

For patient demographics, we extracted mother’s age at delivery, race, ethnicity, and insurance. We selected the most common self-report when there were inconsistencies within a patient’s history of self-reported race or ethnicity. Lab test names and units were mapped to logical observation identifiers names and codes (LOINC). Values were cleaned (invalid results and text removed) and converted to numeric values or standardized to non-numeric ordinal scales depending on the test (e.g., ‘positive’ set to 1 and ‘negative’ set to 0). The earliest result was retained when there were duplicates (e.g., ‘preliminary’ and ‘final’ results from the same test with the same values). Vital signs, including weight, height, temperature, respirations, pulse, oxygen saturation, diastolic blood pressure (DBP), systolic blood pressure (SBP), and self-reported pain were standardized to common names and unit scales. Diagnoses from ICD-9-CM and ICD-10-CM were combined via mapping to single-level diagnosis categories using the Clinical Classifications Software[1]. 96.2% (68,293/70,948) of patients have diagnostic information in the window between 8 months prior to the start of pregnancy up to and including 1 day before delivery. When no timestamp was available for a diagnostic code (only a day), the timestamp was imputed to 11:59pm to avoid any future bias. We filtered medications to those administered (i.e., directly given to patients) and mapped the remaining medication names to RxNorm common ingredients regardless of brand or dose. Procedures were recorded through CompuRecord, an anesthesia information management program, using CPT-4 codes. 17.9% (12,722/70,948) of patients have procedural information in the window between 8 months prior to the start of pregnancy up to and including 1 day before delivery. When procedures included multiple timepoints (e.g., procedure start, anesthesia given, fluid given), only the earliest one was retained.

**Digital phenotype algorithm for postpartum hemorrhage**
Deliveries followed by postpartum hemorrhage were identified using a previously developed, rule-based digital phenotype that was validated against gold standard labels from physician chart review (89% accuracy). The phenotype combines diagnostic and treatment information, which yielded higher accuracy than using blood loss alone to designate PPH status (65-67% accuracy, depending on the cutoff)[24]. Specifically, any of following criteria was sufficient for PPH classification, 1) cumulative blood loss was >1000mL, 2) hematocrit (a proxy measure for blood loss) was critically low (<=21%) or dropped substantially (<=12-point drop from admission baseline to <=25%) within 48 hours of delivery, 3) receipt of PPH-specific medications including carboprost tromethamine, misoprostol, or tranexamic acid within 48 hours of delivery, 4) one of 30 PPH ICD-9/10 diagnostic codes in addition to methylergonovine within 48 hours of delivery (a medication given less specifically for PPH), or 5) any surgical intervention for PPH including Bakri balloon placement, placement of compression sutures and uterine artery ligation or embolization, curettage, or hysterectomies within 48 hours of delivery. For additional details on phenotype development and validation, please see the original publication[24]. Significant differences between cases and controls on demographic and clinical characteristics were determined using chi-square tests for independence (proportional differences) and logistic regressions (continuous measurements).

**Feature engineering**

*Diagnoses, procedures and medications.* For diagnoses, we set a patient’s value to the earliest week they received any ICD-9/10 in a given disease, defined using the approximately 280 clinical classification software (CCS) single-level diagnosis categories[2–4], where week one was eight months prior to pregnancy. For any given diagnosis category, if a patient did not receive a code in that category, their value was set to zero. The same approach was used for procedures (CPT-4 codes) and medications given prior to hospital admission for delivery (any medication brand or dose mapped to
the generic drug). This strategy allowed us to retain initial timing information in addition to presence or absence of the feature.

For medications given during hospital admission for delivery, we quantified the frequency of administration (unique timestamps between admission and delivery times), as well as the total dose for the sixteen most commonly given drugs. We limited dosing variables to the most common for a number of reasons: 1) it was not available for every administration, 2) it required significant additional standardization across units, and 3) variation across patients is limited among medications given less commonly in a relatively short time window (admission to delivery).

Vital signs. For vital signs (excluding pain), we generated summary features for each of the two time periods (prior to and after admission) including minimum, median, and maximum values. We also generated 10 functional principal components using the R package fdapace for each vital sign within each time period using the maximum daily value for each patient when there was more than one. Time was standardized across individuals by taking the difference between the measurement day and the first day of the pregnancy. Functional principal component analyses (fPCA) construct eigenvectors that best represent the covariance matrix where each observation is a time series of values (for example, a patient’s temperature across pregnancy). Individual scores for each fPC were generated for all patients based on fPCAs derived exclusively using the training data. For pain, the maximum value across physical locations for each time period was retained. We did not impute missing values for any vital signs.

Lab values. Numeric lab values were handled similarly to vital signs. We extracted the same summary values, as well as frequency, and generated fPCs as described above. However, since there were hundreds of lab tests, we only derived fPCs for 15 of the most common lab tests and we did not split values by time period because there were too few values post-admission. We additionally
generated baseline values (the value with the earliest timestamp) and maximum change values (the largest shift from baseline during the period) for labs occurring post-admission. For nominal lab values, the maximum ordinal value for each time period was extracted. Missing lab values were left missing.

**Known risk factors.** We manually curated 59 variables spanning demographics, obstetric characteristics, admission baseline vital signs, admission baseline hematocrit and platelet lab values, pregnancy complications, and general medical history that were included in current risk toolkits or used in previously published risk models for inclusion in our feature selection process. The demographic variables we used were age at admission, race (Black, Asian, White, Native American, Other), ethnicity (Hispanic or Latino, Not Hispanic or Latino), insurance (Medicaid or Medicare, private insurance, uninsured, other insurance or missing), alcohol history, and tobacco history. Obstetric characteristics included gestational weeks at admission, dummy variables for gestational weeks <32, between 32 and 37, and >37, delivery method (Cesarean or vaginal), labor trial, spontaneous labor, induced labor (as indicated by oxytocin administration records), parity, a dummy variable for parity >4, prior preterm birth, multiple gestation, and time from admission to delivery. For baseline admission vital signs and labs, we extracted the first available value after admission for SBP, DBP, temperature, height, weight, hematocrit, and platelets. We additionally had dummy variables for baseline hematocrit <30, baseline platelets <100,000, and baseline platelets <70,000. Body-mass index (BMI) at admission, as well as prior to pregnancy, were included, as well as dummy variables for admission BMI >35 and BMI >40. Weight prior to pregnancy was also included.

For pregnancy complications, obstetric history, and medical history, we used binary variables based on ICD codes and CPT-4 procedure codes. The following conditions were included: assisted reproductive technology, breech presentation, chorioamnionitis during hospital admission, early false
labor (<37 weeks), eclampsia, fibroids, group B streptococcal infection, large for gestational age (antenatal diagnosis), preeclampsia (measured using ICDs as well as using a more comprehensive digital phenotype), severe preeclampsia, superimposed preeclampsia, placental abruption, placenta accreta, placenta previa, polyhydramnios, prolonged second stage of labor, premature rupture of membranes (PROM), small for gestational age (antenatal diagnosis), vaginal bleeding, prior miscarriage or abortion, prior PPH, prior stillbirth, anemia, asthma, coagulopathy, depression, diabetes, gastrointestinal disease, gestational diabetes, heart disease, chronic hypertension, renal disease, seizures, and thyroid disease.

**Feature selection**

Among the thousands of features with five or more non-missing values, we used three strategies to select a subset of features for input into our risk model in order to avoid overfitting. First, we used the python package XGBoost to estimate feature importance within type (diagnoses, procedures, medications, vital signs, labs). To get a robust estimate of feature importance, we used the 75th percentile importance score for each feature across 100 bootstrapped (60% randomly sampled per run) gradient boosted decision tree models. We also used adaptive lasso (implemented through the R package caret) and univariate logistic regressions to subset features based on association with PPH. For diagnoses, procedures, and medications, we used coefficients from best adaptive lasso regression in five-fold cross-validation (i.e., lowest mean error) run within each type since these features have no missing values. For vital signs and labs, we used coefficients from univariate logistic regressions for each feature among patients with non-missing values that included race, ethnicity, age, smoking history, alcohol history, insurance type (private, public, other), and gestational weeks at admission as covariates.
Our final feature selection step considered information across these analyses. We find that using multiple methods, relying on both tree-based importance metrics and regression-based p-values, resulting in more robust feature selection. Features with non-zero importance scores from gradient boosting models that also showed significant association with PPH in either lasso or logistic regression (depending on the feature type) were retained. For lasso, we considered an association p-value < 0.01 as significant and for logistic regression, we considered a Bonferroni-corrected p-value < 0.05 adjusting for the total number of vital and lab features to be significant.

**Learning Algorithm**

Selected features were used to train gradient boosted decision tree models with 100-fold cross-validation employed within the training dataset with the python package LightGBM. Gradient boosting trains weak tree models sequentially, using prior errors to inform subsequent models, resulting in stronger performance. LightGBM is a particularly fast and accurate version of gradient boosting that builds trees leaf-wise (by branching just from the best leaf at each node) rather than level-wise. This approach is appropriate for EMR-derived features, given they are often sparse, have missing values, and can contain complex non-linear interactions.

**Model interpretation and simplification**

We used Shapley values to estimate relative importance of each feature using the python package SHAP[5,6]. Shapley values decompose the deviation from expected risk for each individual into the contributions of each feature by measuring the degree to which each feature pushes an individual’s predicted risk above or below the base value, where the base value represents the average predicted risk for the sample. For some patient Y, the Shapley value for feature X is estimated by comparing patient Y’s predicted risk to their predicted risk when feature X is missing. This is actually relatively complex to estimate since a feature’s contribution to a non-linear, interactive model (such as those
based on decision trees) depends on other feature values, so predictions generated from all possible subsets of features must be considered. In practice, Shapley values can be calculated efficiently using TreeExplainer, which takes advantage of the model’s known tree path structure to reduce the number of perturbations that must be evaluated. Once the patient-feature matrix of Shapley values has been generated, overall feature importance can be calculated by taking the mean across individual Shapley values for each feature.

It can also be informative to show dependency plots, which illustrate the relationship between patients’ values for a particular feature of interest and the corresponding increase in predicted risk attributable to that value. This can reveal inflection points above or below which risk is substantially increased or decreased. Since the raw output of LightGBM for classification is in logit (e.g., here, the raw predicted risk is the predicted log odds of PPH), Shapley values for our model were also in logit. To transform them to probabilities, we used a sigmoid function $\sigma$ to compute relative risk (RR) as follows

$$RR = \frac{\sigma(\phi_0 + \phi_i)}{\sigma(\phi_0)}$$

where $\phi_0$ is the base Shapley value and $\phi_i$ is the Shapley value for a feature of interest, reflecting the degree to which that feature increased or decreased predicted risk from the base value for each patient[7].

Towards the goal of building an interpretable, stream-lined model for predicting risk, we tested the performance of our model using only the most important features. We tested our model using only the top feature and then sequentially adding the next important feature across the entire feature set to determine the minimal number of features needed for maximum AUROC. For each feature subset, we performed 10-fold cross-validation in our training data using the same parameters as our main LightGBM models and extracted the same performance metrics. This allowed us to quantify the
marginal performance increase of each additional feature and find the inflection point at which adding more features yielded insufficiently small performance improvements.

Risk factors used in other risk assessment tools

CMQCC, NYSBOH, and AWHONN risk assessment toolkits assign women to low-, medium-, or high-risk based on presence or absence of 16-17 variables[2–4]. The criteria for each toolkit were made available by Kawakita and colleagues and for CMQCC, we also included high-risk variables listed as intrapartum risk factors[5,6]. To implement these classifications here, we used any potential indication of the condition (ICD codes, procedures, medications, lab values, vital values) available prior to delivery.

For intrapartum assessment using the CMQCC, women with placenta previa, placenta accreta, active bleeding upon admission, platelets <100,000, a known coagulopathy, prolonged second stage of labor, prolonged oxytocin administration (>24 hours), magnesium sulfate during admission, or hematocrit <30 combined with any other risk factor (including those for medium risk) were considered high-risk. Women without any of those risk factors, but who have had a prior Cesarean delivery or other uterine surgery, multiple gestation, more than four previous vaginal births, chorioamnionitis upon admission, prior PPH, large uterine fibroids, fetal weight >4 kilograms, or BMI >35 were considered medium-risk. Women without any risk factors were low-risk. NYSBOH labels were very similar to CMQCC, but with the following exceptions: high-risk included platelets < 70,000 (rather than 100,000) and women with two or more medium risk factors. Additionally, hematocrit <30 plus another risk factor, prolonged oxytocin (>24 hours), prolonged second stage of labor, and magnesium sulfate were all medium-risk factors instead of high-risk. Finally, AWHONN had similar high-risk criteria, but several additional medium-risk criteria. Women with placenta previa, placenta accreta, active bleeding upon admission, platelets <100,000, a known coagulopathy, more than one prior PPH,
hematocrit <30 combined with any other risk factor (including those for medium risk), or two or more medium risk factors were considered high-risk. Women without any of those risk factors, but who have had a prior Cesarean delivery or other uterine surgery, multiple gestation, more than four previous vaginal births, chorioamnionitis, prior PPH, large uterine fibroids, prior fetal demise, polyhydramnios, or induction of labor with oxytocin were considered medium-risk. Family history of PPH was also considered a medium risk factor, but could not be extracted from EMR data, so was excluded here.

We also used 55 features employed by Venkatesh and colleagues to predict PPH to train a model in our data using a very similar approach to theirs in order to compare their risk tool to ours[7]. Their features included age at admission, race, insurance, tobacco history, marital status, education status, recreational drug use history, baseline DBP, SBP, temperature, and weight, pre-pregnancy weight and BMI, gestational weeks at admission, multiple gestation, parity, labor trial, spontaneous labor, assisted reproductive technology, breech presentation, chorioamnionitis upon admission, early false labor (<37 weeks), eclampsia, group B streptococcal infection, large for gestational age (antenatal diagnosis), preeclampsia, severe preeclampsia, superimposed preeclampsia, placental abruption, placenta accreta, placenta previa, polyhydramnios, PROM, small for gestational age (antenatal diagnosis), vaginal bleeding, miscarriage or abortion, prior stillbirth, prior preterm birth, prior Cesarean delivery, anemia, asthma, coagulopathy, depression, diabetes, gastrointestinal disease, gestational diabetes, heart disease, chronic hypertension, renal disease, history of seizures, thyroid disease, magnesium sulfate during hospital admission, antenatal steroids, and fetal macrosomia (fetal weight >4 kilograms). We did not have education status on a sufficient percentage of individuals to use, so we excluded this variable.
## Supplementary Results

Table S1. Training performance for machine learning risk assessment tools

| Risk Assessment Tool                  | Phenotype | Sensitivity | Specificity | PPV     | NPV     | AUROC | F1 Score |
|---------------------------------------|-----------|-------------|-------------|---------|---------|-------|----------|
| Integrated Machine Learning (all 80 vars) | PPH       | 0.30 (0.21) | 0.92 (0.07) | 0.27 (0.14) | 0.93 (0.02) | 0.73 (0.03) | 0.34 (0.04) |
| Integrated Machine Learning (top 24 vars) | PPH       | 0.28 (0.22) | 0.92 (0.07) | 0.24 (0.14) | 0.93 (0.02) | 0.73 (0.03) | 0.33 (0.04) |
| Consortium for Safe Labor Study       | PPH       | 0.30 (0.22) | 0.89 (0.10) | 0.21 (0.12) | 0.93 (0.02) | 0.70 (0.04) | 0.30 (0.04) |
| IML model (all 80 vars) EBL >=1000mL |           | 0.38 (0.26) | 0.94 (0.05) | 0.27 (0.15) | 0.96 (0.02) | 0.87 (0.03) | 0.40 (0.03) |
| IML model (top 24 vars) EBL >=1000mL |           | 0.31 (0.29) | 0.94 (0.06) | 0.24 (0.14) | 0.95 (0.02) | 0.86 (0.03) | 0.36 (0.06) |
| Consortium for Safe Labor Study EBL >=1000mL | | 0.27 (0.23) | 0.94 (0.07) | 0.25 (0.17) | 0.95 (0.01) | 0.78 (0.04) | 0.31 (0.07) |

Means and standard deviations over 100-fold cross-validation were reported.
| Category       | Feature                                                                 |
|----------------|-------------------------------------------------------------------------|
| Demographics   | Age at delivery                                                         |
| Demographics   | Insurance: Medicaid or Medicare                                        |
| Demographics   | Race                                                                    |
| Diagnosis      | Advanced maternal age                                                  |
| Diagnosis      | Adverse effects of medical care                                         |
| Diagnosis      | Anemia in pregnancy                                                    |
| Diagnosis      | Cervical insufficiency                                                  |
| Diagnosis      | Deficiency and other anemia                                             |
| Diagnosis      | Fluid and electrolyte disorders                                         |
| Diagnosis      | Forceps delivery                                                       |
| Diagnosis      | Gestational hypertension                                               |
| Diagnosis      | Higher order multiple gestation                                         |
| Diagnosis      | Hypotension                                                            |
| Diagnosis      | Large for gestational age                                               |
| Diagnosis      | Meconium                                                               |
| Diagnosis      | Multiple gestation                                                     |
| Diagnosis      | Nonreassuring fetal status                                              |
| Diagnosis      | Other hematologic conditions                                            |
| Diagnosis      | Placenta accreta                                                       |
| Diagnosis      | Placental abruption                                                    |
| Diagnosis      | Postpartum infections                                                  |
| Diagnosis      | Preeclampsia                                                           |
| Diagnosis      | Preterm labor with preterm delivery                                    |
| Diagnosis      | Severe preeclampsia                                                    |
| Diagnosis      | Shock in delivery                                                      |
| Diagnosis      | Short gestation; low birth weight; and fetal growth retardation        |
| Diagnosis      | Systemic lupus erythematosus and connective tissue disorders           |
| Diagnosis      | Uterine rupture in delivery                                            |
| Diagnosis      | Vaginal lacerations in delivery                                         |
| Diagnosis      | Venous complications, varicose veins                                    |
| Diagnosis      | Viral infection                                                        |
| History        | Anemia                                                                 |

Table S2. All clinical features
| History | Artificial reproductive therapy |
| History | First reported tobacco use |
| History | Prior PPH |
| History | Prior preterm birth |
| Lab     | Basophil % |
| Lab     | Hematocrit |
| Lab     | Hemoglobin |
| Lab     | Mean corpuscular hemoglobin |
| Lab     | Mean corpuscular volume |
| Lab     | Monocyte % |
| Lab     | Neutrophil Count |
| Lab     | Nucleated red blood cells |
| Lab     | Platelets |
| Lab     | Red cell distribution width |
| Lab     | White blood cells |
| Medication | Betamethasone |
| Medication | Bupivacaine |
| Medication | Citric acid |
| Medication | Docusate |
| Medication | Gentamicin |
| Medication | Magnesium |
| Medication | Oxytocin |
| Medication | Prenatal vit no.105-iron 30 mg-folic acid 1.4 mg-dha 300 mg oral pack |
| Medication | Succinylcholine |
| Medication | Sulbactam |
| Medication | Tobramycin |
| Medication | Water |
| Other   | Gestational weeks at delivery |
| Other   | Labor induction |
| Other   | Labor trial |
| Other   | Time from admission to delivery |
| Procedure | Delivery method |
| Procedure | Dilation and curettage |
| Procedure | Exploratory laparotomy |
| Procedure | Hysteroscopy, surgical |
| Procedure | Ligation or transection of fallopian tube(s), abdominal or vaginal approach |
|-----------|--------------------------------------------------------------------------|
| Procedure | Lysis of adhesions (salpingolysis, ovariolysis)                           |
| Procedure | Myomectomy, excision of fibroid tumor(s) of uterus                       |
| Procedure | Pelvic examination under anesthesia (other than local)                   |
| Procedure | Unlisted procedure, maternity care and delivery                           |
| Vital     | BMI at delivery                                                          |
| Vital     | Diastolic blood pressure                                                 |
| Vital     | Pulse                                                                    |
| Vital     | Respirations                                                             |
| Vital     | Systolic blood pressure                                                 |
| Vital     | Temperature                                                              |
| Vital     | Weight                                                                   |

Table S3. Top clinical features include novel and known risk factors

| Rank | Feature                                         | SHAP Importance | Novelty | % Non-PPH | % PPH |
|------|------------------------------------------------|-----------------|---------|-----------|-------|
| 1    | Cesarean delivery                              | 0.119           | known   | 100%      | 100%  |
| 2    | Minimum red blood cells in hosp. admission      | 0.101           | novel   | 71%       | 61%   |
| 3    | Anemia, thalassemia, or other hemoglobinopathies| 0.099           | known   | 100%      | 100%  |
| 4    | Time from admission to delivery                 | 0.09            | known   | 100%      | 100%  |
| 5    | Oxytocin admin. freq. in hosp. admission        | 0.08            | alternate | 96%    | 98%   |
| 6    | Preeclampsia                                   | 0.067           | known   | 100%      | 100%  |
| 7    | Median SBP in hosp. admission                   | 0.061           | known   | 97%       | 98%   |
| 8    | Race: White                                    | 0.044           | known   | 100%      | 100%  |
| 9    | Median pulse in hosp. admission                 | 0.043           | known   | 96%       | 97%   |
| 10   | Anemia complicating pregnancy                  | 0.037           | known   | 100%      | 100%  |
| 11   | Mean corp. hemoglobin freq. in hosp. admission  | 0.033           | novel   | 96%       | 98%   |
|   | Description                                                                 | Value | Type    | Non-PPH (%) | PPH (%)  |
|---|------------------------------------------------------------------------------|-------|---------|-------------|----------|
| 12| Median red blood cell distribution width in hosp. admission                  | 0.032 | novel   | 84%         | 87%      |
| 13| Minimum SBP in hosp. admission                                               | 0.026 | known   | 97%         | 97%      |
| 14| Minimum DBP in hosp. admission                                               | 0.026 | known   | 97%         | 97%      |
| 15| Minimum absolute neutrophils in hosp. admission                              | 0.025 | novel   | 67%         | 55%      |
| 16| Minimum white blood cells in hosp. admission                                 | 0.025 | novel   | 59%         | 50%      |
| 17| Minimum platelets in hosp. admission                                         | 0.025 | known   | 71%         | 61%      |
| 18| Multiple gestation                                                           | 0.023 | known   | 100%        | 100%     |
| 19| Gestational weeks at admission                                               | 0.022 | known   | 100%        | 100%     |
| 20| Temperature in hosp. admission [fPC #9]                                      | 0.021 | alternate | 92%   | 94%      |
| 21| Prior PPH                                                                    | 0.019 | known   | 100%        | 100%     |
| 22| Maximum SBP in hosp. admission                                               | 0.017 | known   | 97%         | 97%      |
| 23| Maximum pulse antepartum                                                     | 0.016 | alternate | 55%  | 64%      |
| 24| Hemoglobin test freq. in hosp. admission                                     | 0.015 | alternate | 96%  | 98%      |
| 25| Bupivacaine total dose in hosp. admission                                    | 0.014 | alternate | 100% | 100%     |
| 26| Minimum hemoglobin in hosp. admission                                        | 0.014 | known   | 96%         | 98%      |
| 27| Maximum platelets in hosp. admission                                         | 0.014 | known   | 71%         | 61%      |
| 28| Magnesium given in hosp. admission                                           | 0.013 | known   | 100%        | 100%     |
| 29| Assisted reproductive technology                                             | 0.013 | known   | 100%        | 100%     |

Novelty was designated ‘known’ if it is currently used in clinical risk assessments or has been previously reported in the literature. ‘Alternate’ signifies that the risk factor is known, but the measure used here is novel. ‘Novel’ means the feature has not been previously reported as a risk factor for PPH. Colors were used to designate concordance. Blue signifies some version of this measure is used in current clinical practice. Yellow signifies some version of this measure has been previously reported as a risk factor in the literature. Red signifies no previous reports or use. The right two columns show the number of non-PPH and PPH deliveries, respectively, with data available for each feature.
Table S4. Linear fixed effects models for case differences in vital signs prior to delivery

| Case Status Estimate | Cluster Standard Error | t   | p          |
|----------------------|------------------------|-----|------------|
| Systolic blood pressure | -5.19                  | 0.26 | -19.95    | 2 x 10^{-16} |
| Diastolic blood pressure | -3.08                  | 0.18 | -16.93    | 2 x 10^{-16} |
| Pulse                | -2.5                   | 0.23 | -10.90    | 2 x 10^{-16} |

Linear fixed effects models were run for each vital sign with case status and hour prior to delivery (set as a categorical variable for hours -13 to -1). Standard errors were clustered by delivery to account for serial correlation across time.
Table S4. Risk categories across risk assessment tools

| Risk Assessment Tool                             | High-Risk | Medium-Risk | Low-Risk  |
|-------------------------------------------------|-----------|-------------|-----------|
| Integrated Machine Learning (all 80 vars)       | 1,423 (10%) | 4,269 (30%) | 8,543 (60%) |
| Consortium for Safe Labor Study (CSLS)           | 1,423 (10%) | 4,269 (30%) | 8,543 (60%) |
| Intrapartum CMQCC                                | 1,862 (13%) | 4,381 (31%) | 7,992 (56%) |
| Intrapartum NYSBOH                               | 1,935 (14%) | 3,937 (28%) | 8,363 (59%) |
| Admission AWHONN                                 | 3,799 (27%) | 7,649 (54%) | 2,787 (20%) |

Number of patients and percent of test set were listed. For the CSLS and our models, high-risk was set to the top 10% of predicted risk, medium-risk was set to the following 30%, and low-risk was assigned to the remaining deliveries. Clinical risk tools (CMQCC, NYSBOH, and AWHONN) exclusively assign category labels (rather than a continuous probability) using presence of absence of risk factors.
Table S5. Test performance across risk assessment tools for EBL phenotype

| Risk Assessment Tool                  | Sensitivity     | Specificity     | PPV      | NPV      | AUROC     |
|--------------------------------------|-----------------|-----------------|----------|----------|-----------|
| Integrated Machine Learning (all 80 vars) | 0.85 [0.83-0.87] | 0.73 [0.72-0.74] | 0.19 [0.17-0.20] | 0.99 [0.98-0.99] | 0.85 [0.84-0.87] |
| Integrated Machine Learning (top 24 vars) | 0.85 [0.83-0.88] | 0.72 [0.71-0.73] | 0.18 [0.17-0.19] | 0.99 [0.98-0.99] | 0.85 [0.84-0.86] |
| Consortium for Safe Labor Study          | 0.70 [0.67-0.72] | 0.71 [0.70-0.71] | 0.15 [0.13-0.16] | 0.97 [0.97-0.97] | 0.77 [0.75-0.78] |
| Intrapartum CMQCC – high risk          | 0.26 [0.22-0.30] | 0.88 [0.87-0.88] | 0.13 [0.11-0.15] | 0.94 [0.94-0.95] | 0.57 [0.55-0.58] |
| Intrapartum CMQCC – medium risk        | 0.75 [0.69-0.80] | 0.57 [0.56-0.58] | 0.11 [0.10-0.12] | 0.97 [0.96-0.97] | 0.66 [0.64-0.67] |
| Intrapartum NYSBOH – high risk         | 0.29 [0.26-0.33] | 0.87 [0.86-0.88] | 0.14 [0.12-0.15] | 0.94 [0.94-0.95] | 0.58 [0.57-0.60] |
| Intrapartum NYSBOH – medium risk       | 0.72 [0.66-0.78] | 0.59 [0.59-0.60] | 0.11 [0.10-0.12] | 0.97 [0.96-0.97] | 0.66 [0.64-0.67] |
| Admission AWHONN – high risk           | 0.45 [0.41-0.50] | 0.74 [0.73-0.75] | 0.11 [0.10-0.12] | 0.95 [0.95-0.95] | 0.60 [0.58-0.61] |
| Admission AWHONN – medium risk         | 0.91 [0.85-0.97] | 0.18 [0.18-0.19] | 0.07 [0.07-0.08] | 0.97 [0.96-0.97] | 0.55 [0.54-0.56] |

Test performance and bootstrapped 95% confidence intervals were reported.
Figure S1: Feature selection.

Figure S1: Increase in test set AUROC as more features are added to the model (in order of importance). Shaded area indicated 95% confidence interval. Black dotted line indicates 29 derived features that contribute to the top feature model.

Figure S2: Data availability

Figure S2: Percentage of patients with percentage of available data for features in the full model (80 variables, red) and the top features models (24 variables, blue).

Figure S3. Model comparison
Figure S3: ROC curves for all of the models. Curves were drawn for the clinical models (CMQCC, NYSBOH, AWHONN) by assigning values of 0, 0.5, and 1 to ‘low’, ‘medium’, and ‘high’ risk patients.
References

1. Agency for Healthcare Research and Quality. HCUP Clinical Classifications Software (CCS) for ICD-9-CM. Healthcare Cost and Utilization Project (HCUP). www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. 2009.

2. Bingham D, Melsop K, Main E. CMQCC Obstetric Hemorrhage Hospital Level Implementation Guide. Palo Alto, CA: 2010.

3. American Congress of Obstetricians and Gynecologists. Maternal safety bundle for obstetric hemorrhage. https://www.acog.org/-/media/project/acog/acogorg/files/forms/districts/smis-ob-hemorrhage-bundle-risk-assessment-id-admin-intrapartum.pdf. 2019.

4. Association of Women’s Health O and NN (AWHONN). Postpartum Hemorrhage Project: A Multi-Hospital Quality Improvement Program. http://www.pphproject.org. 2017.

5. Kawakita T, Mokhtari N, Huang JC, et al. Evaluation of Risk-Assessment Tools for Severe Postpartum Hemorrhage in Women Undergoing Cesarean Delivery. Obstetrics and Gynecology Published Online First: 2019. doi:10.1097/AOG.0000000000003574

6. Gabel K, Lyndon A, Main E, et al. Risk Factor Assessment. https://www.cmqcc.org/content/risk-factor-assessment. 2015.

7. Venkatesh KK, Strauss RA, Grotegut CA, et al. Machine Learning and Statistical Models to Predict Postpartum Hemorrhage. Obstetrics and Gynecology 2020;135:935–44. doi:10.1097/AOG.0000000000003759