Explaining Machine Learning Models for Early Detection of Pregnancy Risk

by

Yuria Utsumi

B.S. Computer Science and Engineering
Massachusetts Institute of Technology, 2021

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Department of Electrical Engineering and Computer Science
August 4, 2022

Certified by . . . .
David Sontag
Professor
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Explaining Machine Learning Models for Early Detection of Pregnancy Risk

by

Yuria Utsumi

Submitted to the Department of Electrical Engineering and Computer Science
on August 4, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Care management programs for high-risk pregnancies aim to detect pregnant women with pregnancy risk factors early so they can receive proper care or preventative treatment. To detect these women, pregnant members are first detected, then they are checked for high risk diagnosis codes or fed into a risk prediction algorithm. Members predicted to be most at risk are outreached and provided guidance on how to manage or monitor symptoms.

In this thesis, we work with the high risk pregnancy care management team at Independence Blue Cross to (1) build a pregnancy identification algorithm to detect pregnant women earlier in their pregnancy, (2) model impactable pregnancy risk factors, and (3) explain these models' predictions. We introduce a new framework for thinking about explainability methods in healthcare – working in assumptions about a prior understanding a clinician may have about the patient and working with high dimensional, redundant data – and we conduct a user study to examine deployability and impact of these algorithms.

Thesis Supervisor: David Sontag
Title: Professor

Acknowledgments

Firstly, I would like to thank my thesis advisor, Prof. David Sontag. Since the time I was enrolled in his course last Spring, I was constantly inspired by his enthusiasm, passion, and effort for everything he does, and the amount of care he puts towards mentoring his students. I am grateful for his guidance and kindness, and for making this a fulfilling MEng experience.

I am also appreciative of the rest of the Clinical Machine Learning group, for being a kind and supportive group to be around during my MEng. I would like to thank Hussein Mozannar for his mentorship this past year. Thank you for all of your help, from understanding how to formalize problems to planning out phases of the project. I appreciate your patience, support, and thoughtful advice throughout the year. I would also like to thank my co-mentor Irene Chen. Thank you for patiently guiding me through various roadblocks; your energy and enthusiasm was uplifting to be around. Lastly, thank you to Christina Ji for your generosity with answering various IBC-related questions, for being a great desk mate, and for never failing to check in about grabbing lunch in the office.

I would also like to thank our collaborators from Independence Blue Cross (IBC) – in particular, Stephanie Gervasi and Kyle Armstrong, for your help with making this project possible. Thank you, Stephanie, for your endless expertise on the inner workings of IBC and for leading many of our discussions on formulating this project. And thank you to Kyle for your help with replicating various components of the high risk pregnancy pipeline on our servers and on other technical aspects of the project.

I am also grateful for the high risk pregnancy (HRP) care management nurses for answering our questions about the program. In particular, thank you to Michele N. Ewing for meeting with us throughout the year and helping us make the user study possible, and the rest of the HRP nurses – Patricia Harrigan and Kimberley Berg.

Finally, I would like to thank my family for their constant support throughout my life and my time at MIT. I don't know where I would be without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we introduce the goal of this thesis: explaining machine learning models to detect pregnancy risk early to guide clinical interventions and preventative care. We first outline the clinical context – care management programs for high risk pregnancies – and how machine learning models are used to enroll members. We then argue the need for methods that explain model predictions. We highlight the challenge of explaining model predictions in clinical settings. We conclude the chapter with the structure and scope of the thesis.

1.1 Care Management for High Risk Pregnancy

Care management programs support members with complex help needs to control healthcare cost and/or resource usage. A member identified for care is matched to a care management practitioner who assesses the member's risks and needs, develops a care plan, teaches the member how to manage the disease/symptom, and follows up to track their well being.

High risk pregnancy refers to pregnancies that have potential complications for the pregnant person and/or the fetus. Care management programs for high risk pregnancies aim to detect pregnant members with (potential) pregnancy risk factors early so they can receive proper care, preventative screening, treatment, and educational resources.

One of the several ways that nurse care managers prioritize outreach to pregnant people is through model-based methods. A machine learning model takes, e.g. a pregnant member's medical history (insurance claims and other data domains), as input to predict the likelihood they will have a high risk pregnancy. Members predicted to be most at risk are reached out to and provided guidance on how to manage or monitor symptoms and reduce the potential for adverse outcomes.

1.2 Explaining Model Predictions in Clinical Settings

When deploying machine learning models in the clinical setting, it becomes important to consider how we might provide reasoning for the model predictions to buy in clinicians' trust and help them make informed decisions by explaining model predictions and calling to attention patterns that might otherwise be out of immediate view. Existing works in explaining machine learning models have encompassed intrinsically interpretable models (model-specific methods) to methods that explain any class of models post-hoc (model-agnostic methods), correlation-based methods to causal attribution methods, and methods that explain the model globally to methods that explain individual predictions. Within healthcare applications, past works have studied off-the-shelf methods and interpretable models across different applications [52, 31, 36] and emphasized the importance of evaluations involving a practitioner's perspective and assessing joint human-AI performance [42], though there are various challenges specific to healthcare data that have not yet been addressed in explainability literature.

1.3 Thesis Scope and Organization

In Chapter 2, we start by providing further clinical background on the high risk pregnancy problem and a summary of related works on identifying pregnancies and explainability methods for machine learning. In Chapter 3, we present our algorithm for identifying pregnant members and evaluate how much earlier we are able to detect pregnancy using our hybrid algorithm, compared to an existing non machine learning-based approach. In Chapter 4, we describe the process of building and evaluating a model for impactable risk factors of pregnancy – from defining and validating the target outcomes to analyzing its performance across time and across racial groups. In Chapter 5, we motivate requirements for explainability approaches in clinical applications and present our baselines and methods. In Chapter 6, we present two user studies that examine deployability and impact of (1) pregnancy identification algorithm, and (2) pregnancy risk factor model with and without explanations. Finally, in Chapter 7, we conclude with some remarks and avenues for future work.

Chapter 2

Background

In this chapter, we provide further background on the clinical context of our work: high risk pregnancies & their risk factors, how risk for such pregnancies can be reduced through preventative care, and how care management programs deliver that care by identifying and enrolling high risk members. Finally, we survey related work on early identification of pregnancy, explaining machine learning models, and challenges of explainability in healthcare applications.

2.1 Pregnancy Terminology

We first review common terminology in high risk pregnancy literature.

Gestation. The period of time between conception and birth. Average length of gestation is 40 weeks for humans. *T weeks gestation* means that it has been T weeks since conception. *Gestational* means to occur in or to onset during gestation.

Diabetes Mellitus. A chronic health condition in which the pancreas produces little to no insulin or the insulin cannot properly regulate glucose in the blood due to genetics (type 1) or life style (type 2). *Gestational diabetes mellitus (GDM)* refers to diabetes developed during gestation in members without prior history of diabetes.

Hypertension. Elevated blood pressure characterized by systolic blood pressure $\geq$ 140 mmHg and diastolic blood pressure $\geq$ 90 mmHg. *Gestational hypertension* refers to hypertension that develops during gestation; it may develop into pre-eclampsia.

Pre-eclampsia. Pregnancy complication that occurs after 20 weeks gestation and is characterized by high blood pressure (hypertension) and high levels of protein in the urine. Exacerbations can lead to eclampsia, which has more severe and life threatening outcomes associated.

2.2 High risk pregnancy and early interventions

High risk pregnancy is any pregnancy that has potential complications for the pregnant person and/or the fetus. Pregnancy complications like gestational diabetes and pre-eclampsia can lead to childbirth complications such as eclampsia, cardiomyopathy, and embolism, which can result in adverse pregnancy outcomes such as preterm birth. In 2018, pregnancy and childbirth complications affected 19.6% and 1.7% of pregnancies, respectively, in the U.S. [4]. Pregnant people who have complications during pregnancy are twice as likely to have childbirth complications than those who did not have complications [4].

Risk factors for complications include advanced or young maternal age, maternal health problems, multiple pregnancy, and pre-existing health conditions (e.g. diabetes, high blood pressure) [29]. With the rise in people entering pregnancy with pre-existing conditions, the number of pregnancy and childbirth complications have consequently increased in the past few years [4]. Additionally, there exist systemic disparities in pregnancy and childbirth complications, e.g. Black people have mortality rates over three times higher than White people during pregnancy [44], which makes this a crucial problem from a fairness perspective.

Often times, risk factors for adverse outcomes can be addressed to reduce risk for adverse outcomes downstream. Frequent antenatal care visits lead to lower maternal, fetal, and neonatal morbidity and mortality [32]. For pregnant people with gestational hypertension, anti-platelet agents reduce pre-eclampsia onset by 40%; calcium supplements reduce risk of pre-eclampsia by 50-60% [32]. Physical activity reduces risk and development of hypertension [21]. For pregnant people with gestational diabetes, adopting a low glycemic index diet and increasing physical activity help reduce glucose levels, which in turn reduces fetal macrosomia [1] and weight gain [13].

Though existing research is inconclusive, *early* intervention and treatment can be particularly important in reducing risk for adverse outcomes. Increased physical activity from pre-gestation, up to 20 weeks gestation reduces risk for gestational diabetes [55], and early treatment for gestational diabetes (before 24 weeks gestation) reduces rates for large for gestational age infants (p=0.03) [46] and decreases rates for pre-eclampsia (p=0.03) [48]. Intervention to control glycemic levels early in pregnancy reduces risk of pre-eclampsia in people with type 1 diabetes [26], and treatment of early onset mild gestational hypertension before 20 weeks gestation reduces maternal and fetal complications [58]. Other studies show no beneficial effect of early diagnosis and treatment [8, 53, 28].

Furthermore, high risk pregnancies generate some of the highest costs and resource usage in medicine [12], with childbirth related hospitalizations generating $16 billion or 4% of in-patient hospital costs in the U.S. in 2008 and pregnancy and childbirth accounting for almost 25% of hospitalizations in the U.S. in 2003 [45]. Thus, parties that incur the costs of care and resource, like insurance companies, are highly incentivized to identify high risk members and deliver care early on in their pregnancy to decrease future complications, care, and consequently costs.

2.3 Pregnancy care management

To deliver these preventative cares, care management programs for high risk pregnancies identify and enroll high risk, high cost members early in their pregnancy to proactively manage their health and control cost [27, 39, 7]. After pregnant members are identified, the high risk, high cost members are identified via predictive models

[1] Development of a fetus that is larger than the average size of 4000-4500 grams.

and rule-based methods (e.g. member meets a certain diagnosis criteria) [11, 39]. Once identified, care managers (e.g. nurses) educate expecting members about their disease, condition, or medication and care for risky members [39]. For this thesis, we work with maternity program nurse care managers at Independence Blue Cross (IBC)[2]. Specifically, we work with the enrollment pipeline for this program, which uses both clinical codes and predictive machine learning models to prioritize members for care outreach based on complex health needs related to pregnancy.

2.4 Related Work

In this section, we discuss prior work relevant to the early pregnancy identification algorithm we develop, as well as recent work on explainability methods for machine learning models and discussions surrounding challenges in explaining models deployed in healthcare.

2.4.1 Identification of Gestational Episodes

A gestational episode starts at conception and is marked by the first day of the pregnant person's last menstrual cycle, and the end is marked by the pregnancy outcome event [49]. Given a member's medical history, gestational episodes can be inferred retrospectively or identified in an online manner. Much of the existing literature covers the former [38, 10, 37, 6, 49]; the latter is more difficult since the last menstruation date is generally not recorded in insurance claims data [49].

Some retrospective methods infer gestational episodes by identifying the pregnancy outcome, then subtracting an estimated gestational age for the start of gestation [10, 37, 6, 49]; other methods additionally use a set of early pregnancy markers, such as positive urine pregnancy test or nuchal ultrasound, for estimating start of gestation [38]. Moreover, [38] works with claims databases under the Observational Medical Outcomes Partnership (OMOP) Common Data Model, which is the format

[2] Independence Blue Cross is a health insurer based in Philadelphia, Pennsylvania and the largest health insurer in the Philadephia region.

in which the IBC data is stored.

In this thesis, we develop an algorithm that identifies gestation episodes in an online manner. We build on the set of gestation start and outcome markers in [38] to develop a hybrid model-based algorithm that performs at least as well as identifying episodes from the codes alone.

2.4.2 Explaining Machine Learning Models

Explainability is described as "meaningful information about the logic behind automated decisions using their data." [3]. Traditionally, explainability methods are categorized into model-specific and model-agnostic methods. Model-specific methods are models that are intrinsically interpretable, such as decision trees, rule-based models, linear regression, and attention networks [54, 35, 47, 59]. Model-agnostic methods explain model predictions in a post-hoc fashion by constructing interpretable local surrogate models [47, 18, 34], computing gradients to assess how changes in input affect model predictions [51, 50], and scoring the importance of covariates on the prediction [36, 23]. Broadly, these methods try to attribute responsibility for model output to the model inputs (i.e. covariates) and surface covariates that are most responsible for the prediction.

However, many of these attribution methods are correlation-based and therefore fail to provide causal attributions, i.e. attribute responsibility to covariates causing the prediction, especially in the presence of confounders. Causal explanations are important, particularly in high stakes applications (like healthcare), where such explanations are key to establishing user trust and informing clinical decision-making [30, 52]. As a result, recent literature has started to shift towards causal attribution methods [14, 41, 30, 19], though existing methods must make many simplifying assumptions about the variables' causal relationships to circumvent the difficulties of inferring the structural causal model from data. Therefore, a generalized, versatile method, e.g. a causal version of SHAP or LIME, does not yet exist.

Explainability in Healthcare

In healthcare, explanations serve the purpose of gaining clinician's trust and helping them make informed decisions by working collaboratively with the AI. Within this area, existing works have focused on applying simple interpretable models like regression and naive Bayes models [52], developing intrinsically interpretable models [31], developing techniques for different data modalities [15, 33], and comparing off-the-shelf explainability algorithms [17, 36]. Since practitioners often interact with these models in member-specific scenarios, there is also a large focus on local explainability methods which aim to explain a machine learning model's prediction at the member level [36, 9].

Moreover, there is a need for evaluation frameworks for explanation methods as they are rarely tested [22]. While some existing methods systematically evaluate explanations by e.g. measuring concordance among different methods or alignment with domain knowledge [5], evaluation from a human perspective is important to check for correctness since models (and surrogate models) may not fully capture the true workings of the system being modeled [22]. Furthermore, it is important to evaluate explanations in a user-centered environment to test for human-centered principles like whether the explanation enhances joint human-AI performance or whether the explanation helps the user distinguish between trustworthy and untrustworthy model outputs [42].

However, not much has been proposed in terms of how to formalize explainability in a clinical setting – in particular, how to address challenges and scenarios specific to healthcare data.

In this thesis, we introduce a new framework for thinking about explainability methods in healthcare – in particular, working in assumptions about a prior understanding a clinician may have about the member and working with high dimensional, redundant data. We also conduct a user study to evaluate explanations when working jointly with a nurse for a high risk pregnancy care management program.

Chapter 3

Early Identification of Pregnancy

Pregnancy identification is a task that identifies whether or not a member is in gestation. In this chapter, we discuss a framework for identifying pregnant members from their medical history in an online manner. We build on the set of clinical codes that indicate gestation start and pregnancy outcomes [38] to develop a hybrid machine learning-based algorithm that performs at least as well as identifying the episodes from clinical codes alone.

3.1 Cohort Selection and Dataset Generation

We begin by selecting a cohort of pregnant and non-pregnant members and feature set for the model. We use terminology consistent with the OMOP Common Data Model (OMOP CDM) [2], which is the form in which the data is stored.

3.1.1 Cohort Selection

For evaluation purposes, we divide the members into 3 sub-cohorts as follows.

1. **Pregnancies without complications.** Pregnant members whose most recent pregnancy had a live birth outcome, but no adverse outcome or complications in the same gestational episode.

2. **Pregnancies with complications.** Pregnant members whose most recent pregnancy had an adverse outcome or complication, e.g. hypertension/pre-eclampsia, ectopic pregnancy, neonatal ICU.

3. **Never pregnant members.** Female members of child-bearing age who never entered gestation.

We build the pregnant sub-cohort using the the algorithm in [38]. We modify the algorithm to consider other (adverse) pregnancy outcomes, such as onset of pre-eclampsia and newborn admission to the neonatal ICU. The algorithm infers the start and end of the most recent pregnancy episode, and the corresponding pregnancy outcome or complication. The full algorithm is defined in A.1.1. We describe the high level algorithm in Figure 3-1. Note that members in the no complications sub-cohort are those returned by the algorithm with a "live birth" outcome; all remaining members belong to the sub-cohort with complications.

We build the never pregnant sub-cohort by sampling female members who never have codes indicating gestation start or pregnancy outcome, according to the age distribution of the pregnant sub-cohort. The full algorithm is defined in A.1.2.

This gives us an overall cohort of 2,397,956 members (27.7% pregnancies without complications, 53.9% pregnancies with complications, 18.4% never pregnant), with average age 31.8 years (4.8 std), 32.7 years (6.7 std), and 32.1 years (6.1 std), for pregnancies without complication, pregnancies with complication, and never pregnant sub-cohorts, respectively.

3.1.2 Dataset Generation

We use members' medical history to construct features for the pregnancy identification model. We derive pregnancy labels using the dates of pregnancy start ($t^{'}_{start}$) and outcome codes ($t^{'}_{end}$). The process for each sub-cohort is as follows (see Figure 3-2 for an illustration):

1. **Pregnancies without complications.** We assume full term pregnancy and set pregnancy start to be 40 weeks prior to the outcome date, i.e. t_{start} =

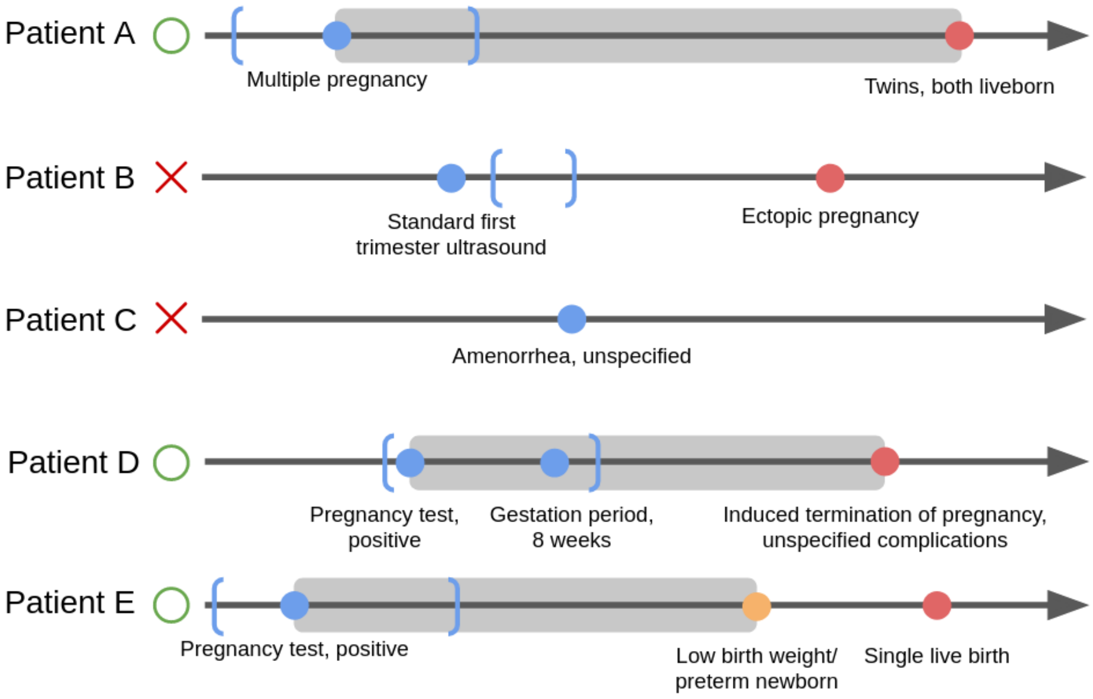


Figure 3-1: Illustration of the pregnancy cohort selection algorithm (3). First, the most recent pregnancy outcome is detected (red point). Then, we search for pregnancy start code(s) (blue point(s)) within a specified lookback window for the corresponding outcome (blue brackets); the earliest start code marks the start of that pregnancy episode. Finally, we do a forward search for any additional pregnancy outcome or complications (orange point); if one exists, the pregnancy outcome is updated.

Member B is excluded from the cohort since no pregnancy start code was detected within the lookback window. Member C is excluded since there was no associated pregnancy outcome code; amenorrhea alone cannot indicate pregnancy has started since it can be caused by non pregnancy-related factors (e.g. stress, menopause).

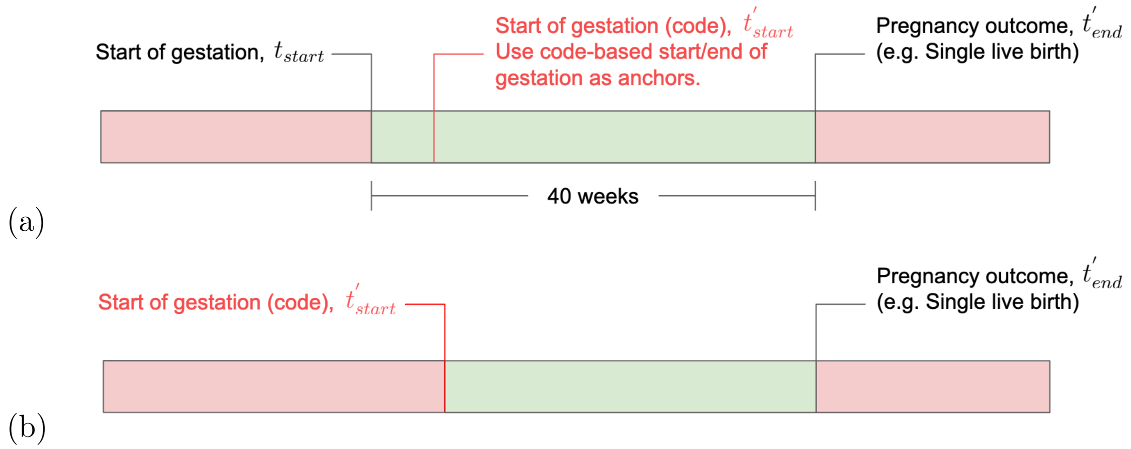

Start of gestation, t_{start}
Start of gestation (code), $t^{'}_{start}$
Use code-based start/end of gestation as anchors.
Pregnancy outcome, $t^{'}_{end}$
(e.g. Single live birth)
40 weeks
(a)
Start of gestation (code), $t^{'}_{start}$
Pregnancy outcome, $t^{'}_{end}$
(e.g. Single live birth)
(b)

Figure 3-2: (a) For pregnancies without complications, we set start of gestation to be 40 weeks prior to when the outcome code is observed, assuming a full term pregnancy. (b) For pregnancies with complications, we set start of gestation to be the date of pregnancy start code.

$t^{'}_{end} - 40$ weeks. If the outcome falls between t_{start} and $t^{'}_{end}$, it has a positive label; otherwise, it has a negative label.

2. **Pregnancies with complications.** Since pregnancies with complications vary in duration, we use the date of pregnancy start code as a noisy pregnancy start date, i.e. $t_{start} = t^{'}_{start}$. Label derivation is the same as above.

3. **Never pregnant members.** All outcomes have a negative label.

For pregnant members, we sample data from 20 weeks before pregnancy start to 20 weeks after the pregnancy outcome is observed (approximately 80 weeks, assuming 40 week gestation period), to allow for early pregnancy and non-pregnancy indicators to be learned, while avoiding signal from previous pregnancies. Data is sampled once a week. For never pregnant members, we sample 80 weeks of data, around the midpoint of their medical history.

For each data point, we generate non-temporal and temporal features from medical data. For temporal data, we construct windowed features, which aggregates the data within a specified backward time window maps them to a binary indicator feature indicating whether the data occurred or not during that time window. Windowed

	Accuracy	**AUROC**	**F1**
Pregnancies without complications	0.8821	0.9454	0.8839
Pregnancies with complications	0.8622	0.9474	0.8641
Never pregnant	0.9811	N/A	0.9990

Table 3.1: Evaluation metrics for f^*, hybrid predictor for pregnancy identification.

features for 5 day and 10 day windows are generated using `omop-learn`[1], a Python machine learning package for OMOP CDM, for the following categories: medical conditions, prescriptions, procedures, specialty visits, and labs[2]. We also include 12 non-temporal features, which include age, race, and gender. This gives us a feature set of 62,734 features. The outcome is a binary label indicating whether the member is pregnant at that time.

3.2 Training Pregnancy Classifier

For each sub-cohort, we split the data into train (50%), validation (25%), and test (25%). We aggregate all three sub-cohorts to construct the train and validation sets and learn a function $f : R^k \to \{0, 1\}$, mapping from features to a binary label (pregnant, not pregnant). We fit LASSO (L1-regularized) logistic regression and select model f^* with the highest validation accuracy. See A.3.1 for training configurations.

3.2.1 Evaluation

We compute evaluation metrics using a hybrid predictor, i.e. pregnancy start codes and outcome codes are used as anchors [24], or weak labels for pregnancy outcome. We report accuracy, AUROC, and F1 score on the test set, in Table 3.1.

3.3 Inferring Pregnancy Episodes

Once the model f is fit on training data, we infer pregnancy episodes (pregnancy start and end) for each member in the test set. The hybrid algorithm smooths the predicted

[1] `https://github.com/clinicalml/omop-learn`

[2] We remove pregnancy start codes and outcome codes, which are used as anchors. [24]

model probabilities and sets start of pregnancy to be the earlier of: pregnancy start code or first indication of positive prediction with increasing trajectory. Then, end of pregnancy is inferred similarly. We outline the procedure below in Algorithm 1 and 2.

For pregnant members, we infer pregnancy episodes under two settings: with nurse filtering and without nurse filtering. Nurse filtering simulates what would happen if a nurse filtered for pregnancy triggers that are too early. We define "too early" to be any pregnancy start triggers that occur before a month after the true start of pregnancy, since we don't expect pregnancy-related codes to appear in this window.

Algorithm 1 Inferring pregnancy start and end for each member.

```
for $i \in P$ do
    $\hat{p} \leftarrow f(X_i)$                          ▷ predict probability of pregnancy over time
    $\hat{q} \leftarrow$EMA($\hat{p}$)                   ▷ smooth with exponential moving average filter
    $\hat{y} \leftarrow$ predict($\hat{q}$)              ▷ returns binary predictions
    $\hat{start}_i, \hat{end}_i \leftarrow$InferEpisode($\hat{q}, \hat{y}$) ▷ infer pregnancy start and end (see Alg. 2)
end for
```

3.3.1 Evaluation

We evaluate how much the algorithm helps with identifying pregnant members earlier and characterize the members who are identified pregnant by the model. We also evaluate false positives on non-pregnant members.

Comparing pregnancy identification algorithm to the code baseline. For the pregnant sub-cohorts, we plot a distribution of δ_i, how much later pregnancy was detected, past t_{start} (pregnancy start date) for member i. For pregnancies with complication, we evaluate on a subset who have live birth outcomes and set $t_{start} = t'_{end,LB} - 40$ weeks, where $t'_{end,LB}$ is the date of the live birth outcome and we assume they carry to full term. We report the distributions, and metrics comparing the distribution of δ_i's from the code baseline (blue) to the algorithm (pink) in Figures 3-3 and 3-4 for pregnancies without and with complication, respectively.

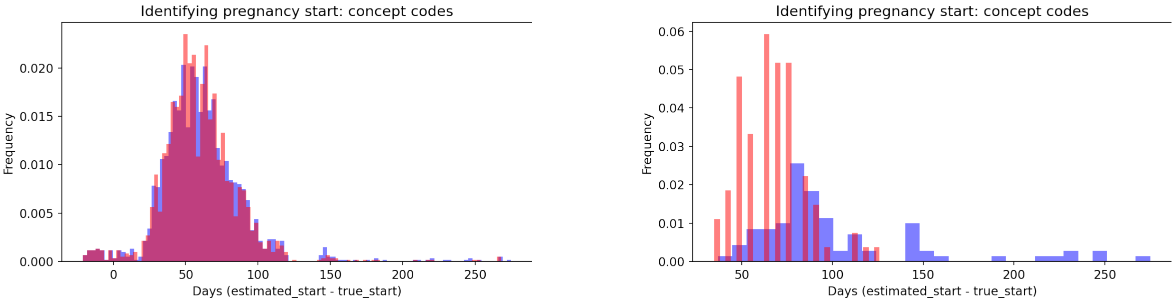
[3] Simulated nurse filtering filters out model predictions that occur too early for pregnancy signal to exist (i.e. before 1 month after t_{start}).

Algorithm 2 Inferring pregnancy start and end, given smoothed probability and predictions over time $(\hat{q}, \hat{y})$.

```
isStart=True; $l$=len($\hat{q}$)
start, end = None, None
for $t = 0 : l - 2$ do
    if (isStart) and ($\hat{y}[t]$ == 1) and ($\hat{q}[t] < \hat{q}[t+1]$) then
        // set pregnancy start if we have +ve prediction and increasing probability
        start← $t + 1$; isStart←False
    else if (not isStart) and ($\hat{y}[t]$ == 0) and ($\hat{q}[t] > \hat{q}[t+1]$)) then
        // set pregnancy end if we have -ve prediction and decreasing probability
        end← $t + 1$
    end if

    // use code-based prediction by default if pregnancy start is before
    // 1 month after true pregnancy start (and we are simulating nurses filtering)
    if nurseFilter and start < trueStart+deltaMonth then start←codeStart
    end if

    // set start and end to be the earliest value (code-based or model-based)
    start← min(codeStart, start)
    end← min(codeEnd, end)
end for
```



Full distribution:

Code distribution

84.0/91.0/103.0 days (85th/90th/95th %le)

Algorithm distribution

82.0/89.0/98.9 days (85th/90th/95th %le)

Distribution of model δ_i's (4.29%):

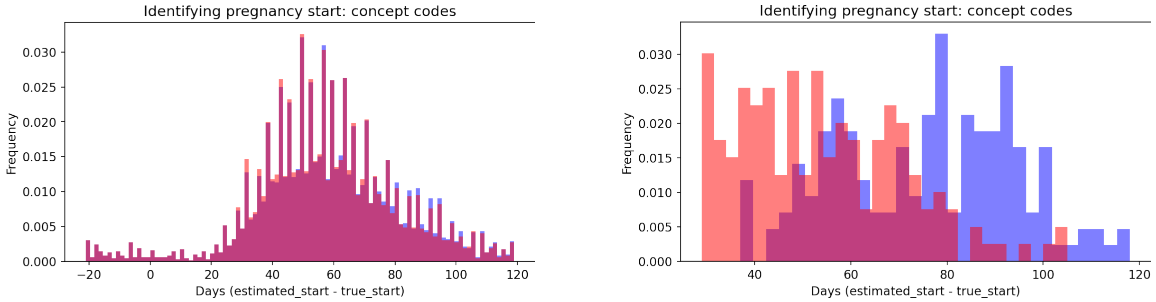
Code distribution

145.0/152.4/226.4 days (85th/90th/95th %le)

Algorithm distribution

84.0/85.4/95.2 days (85th/90th/95th %le)

Figure 3-3: Distribution of pregnancy identification delay for pregnancies without complication, simulating nurse filtering[3]. (left) shows the full distribution; (right) shows the distribution of members who were identified early by the model (4.29% of members).



Full distribution:

Code distribution

83.0/90.0/98.0 days (85th/90th/95th %le)

Algorithm distribution

81.0/88.0/98.0 days (85th/90th/95th %le)

Distribution of model δ_i's (3.54%):

Code distribution

94.6/99.4/103.8 days (85th/90th/95th %le)

Algorithm distribution

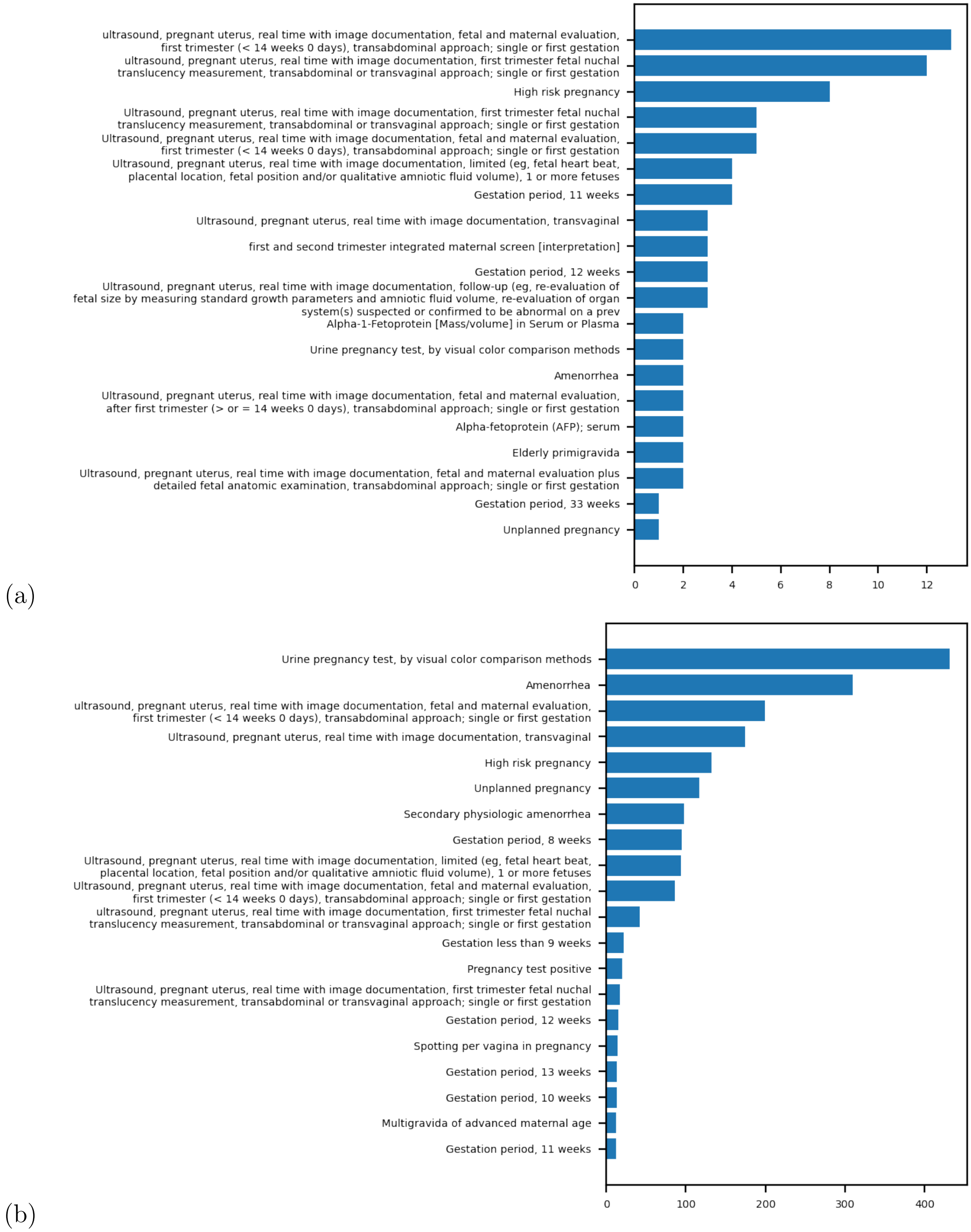
71.6/77.0/82.8 days (85th/90th/95th %le)

Figure 3-4: Distribution of pregnancy identification delay for pregnancies with complication, simulating nurse filtering[4]. (left) shows the full distribution; (right) shows the distribution of members who were identified early by the model (3.54% of members).

Characterizing members predicted early by the model. We compare the distribution of pregnancy start codes for members detected pregnant earlier than the code by the model to members who are detected by the code in Figures 3-5 and 3-6 for pregnancies without and with complications, respectively. From these plots, we can see that codes for high risk pregnancies and ultrasounds are more prevalent in the group detected by the model. Codes for urine pregnancy test is less prevalent. The model is identifying members who may have started pregnancy visits later in their term since, e.g. they tested for pregnancy using at home tests. This could be reason to e.g. offer cost-free pregnancy tests at local clinics so members are incentivized to get tested formally, and in turn the insurance company obtain data to identify pregnant members earlier. A large proportion of these members also tend to be high risk, which is exactly who we want to identify early for early intervention and treatment.

False positives on non-pregnant members. On the non-pregnant subcohort, we identify pregnancy in 5.58% of members. We report top features surfaced by the model in this subcohort in Table 3.2. Most features are routine labs and procedures

[4] Simulated nurse filtering filters out model predictions that occur too early for pregnancy signal to exist (i.e. before 1 month after t_{start}).



(a)

(b)

Figure 3-5: Comparison of top pregnancy start codes for (a) members detected pregnant by the model versus (b) members detected pregnant by codes, for pregnant members without complications with simulated nurse filtering.

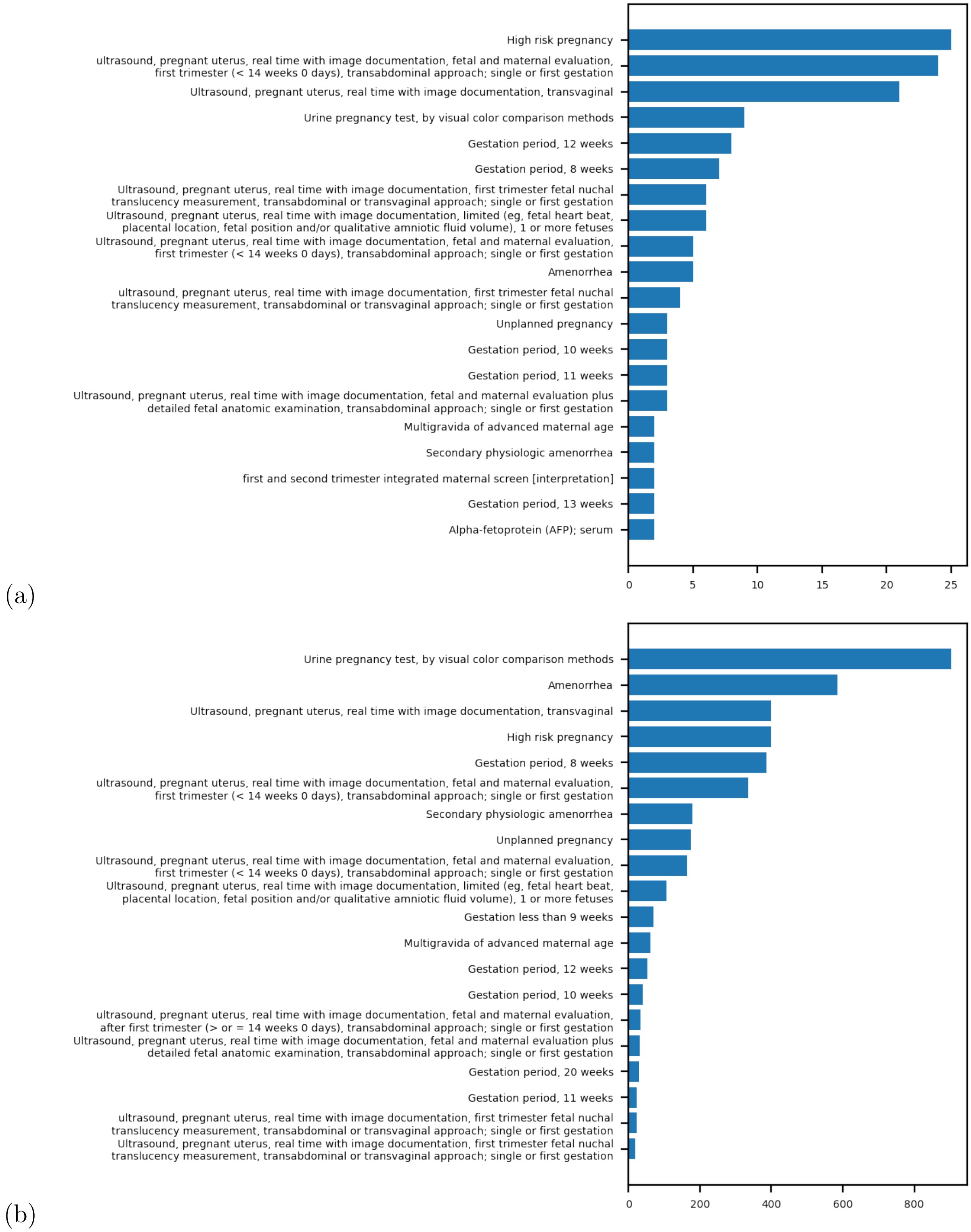


Figure 3-6: Comparison of top pregnancy start codes for (a) members detected pregnant by the model versus (b) members detected pregnant by codes, for pregnant members without complications with simulated nurse filtering.

Feature name
2213418 - procedure - Immunization administration (includes percutaneous, intradermal, subcutaneous, or intramuscular injections); 1 vaccine (single or combination vaccine/toxoid)
2212167 - labs - Urinalysis, by dip stick or tablet reagent for bilirubin, glucose, hemoglobin, ketones, leukocytes, nitrite, pH, protein, specific gravity, urobilinogen, any number of these constituents; non-automated, without microscopy
2108115 - procedure - Collection of venous blood by venipuncture
3050479 - labs - Immature granulocytes/100 leukocytes in Blood
2212996 - labs - Culture, bacterial; quantitative colony count, urine
3033575 - labs - Monocytes [/volume] in Blood by Automated count
3023314 - labs - Hematocrit [Volume Fraction] of Blood by Automated count
3014576 - labs - Chloride [Moles/volume] in Serum or Plasma
38004461 - specialty - Obstetrics/Gynecology
3015746 - labs - Specimen source identified

Table 3.2: Top positive features surfaced by non-pregnant members who were inferred to be pregnant.

performed on pregnant members, but doesn't necessarily imply pregnancy. Similarly, a specialty visit to an obstetrician or gynecologist can be made outside of pregnancy, e.g. for fertility or menstruation issues.

Chapter 4

Modeling Pregnancy Risk Factors

In this chapter, we detail the process of building and evaluating a model for impactable pregnancy risk factors. We describe how the target outcomes are defined, and how we use those definitions to generate the cohort and dataset. Once trained, we evaluate the model to show how it performs at different points in pregnancy and how early it catches risk. We also audit the model for racial fairness analysis.

4.1 Defining the Target Outcomes

For this study, we were interested in identifying members at risk of *impactable* risk factors of high risk pregnancy, particularly gestational diabetes and hypertension / pre-eclampsia. Though differentiating between gestational hypertension and chronic hypertension does not help with establishing pregnancy risk [57], we model gestational hypertension (hypertension onset during pregnancy) since chronic hypertension can be inferred simply by querying codes in the member's history.

To compile codes for these outcomes, we queried for pregnancy episodes with a gestational diabetes ICD 10 code (O24.11-O24.93) using ATLAS[1], then filtered for unique diagnosis codes within those episodes. We selected the most frequently occurring diagnosis codes as the initial set of target codes for gestational diabetes outcome.

[1] ATLAS is a software by the Observational Health Data Sciences and Informatics (OHDSI) community that can be used to search and navigate the vocabulary within the OMOP Common Data Model and curate cohort definitions.

	Count (cohort)	**Percentage (cohort)**	**Percentage (literature)**
Pregnancy episodes	57,183	–	–
Pregnancy episodes w/ gestational DB code	3,922	6.86%	6.5-11.9% [61] 7.5-8.9% [60]
Pregnancy episodes w/ gestational HT/PE code	5,394	9.43%	1.8-4.4% (gHT) 0.2-9.2% (PE) [56]

Table 4.1: Comparison of gestational diabetes and gestational hypertensive members' prevalence in cohort versus literature.

The same procedure was repeated for gestational hypertension / pre-eclampsia (ICD 10 code O10.011-O16.9).

To validate, we queried for pregnancy episodes containing the target codes for each of the outcomes to compare against their prevalence numbers from literature. We show a summary in Table 4.1.

We additionally validated the code set with the care management nurses, who hand-labeled outcome codes for a subset of 20 members, given data up to the end of pregnancy episode. This allowed us to (1) validate that the existing codes are indicative of the corresponding outcome, and (2) find new codes indicative of an outcome. Methyldopa 250 MG Oral Tablet, an anti-hypertensive drug, was added as a code for gestational HT/PE.

4.2 Cohort Selection and Dataset Generation

4.2.1 Cohort Selection

To build a cohort of pregnant members, including ones with gestational diabetes and hypertension complications, we again referred to our pregnancy cohort selection algorithm (3). We select the subset of members with outcomes: live birth (no complication), gestational hypertension, and gestational diabetes. This gives us an overall cohort of 12,243 members.

	Accuracy		**AUROC**	
	Mean	**90% CI**	**Mean**	**90% CI**
LASSO (L1)	0.7678	0.7633-0.7722	0.7606	0.7555-0.7656
ELASTIC-NET (L1+L2)	0.7129	0.7082-0.7177	0.7357	0.7304-0.7410
XGBOOST	0.6866	0.6818-0.6915	0.7696	0.7648-0.7745

Table 4.2: Evaluation metrics for g^*, hybrid predictor for pregnancy risk factors.

4.2.2 Dataset Generation

We use the medical data to construct features for the risk model. To ensure our data is properly distributed across different stages of pregnancy, we sample 10 data points for each member, uniformly distributed across the following time slices: 3 months before pregnancy start, trimester 1, trimester 2, and trimester 3.

Similar to pregnancy identification, we generate non-temporal and temporal features for each sampled point. For temporal data, we generate windowed features for 30 day, 180 day, 365 day, 730 day, and 10k day windows using `omop-learn`[2] for the following categories: medical conditions, prescriptions, procedures, specialty visits, and labs. We also include 12 non-temporal features, which include age, race, and gender. This gives us a feature set of 112,322 features.

4.3 Training

We split the data into train (60%), validation (20%), and test (20%). We learn a function $g : R^d \to \{0, 1, 2\}$, mapping from features to a ternary label (gestational diabetes, gestational hypertension, no complication). We fit several standard classification algorithms – LASSO (L1-regularized), ELASTIC-NET (L1 and L2-regularized), and XGBOOST (gradient-boosted tree)[3] – and select g^* to be the LASSO model with the highest validation accuracy. We report evaluation metrics on the test set in Table 4.2. See A.3.2 for training configurations and A.4.2 for confusion matrices.

[2] `https://github.com/clinicalml/omop-learn`

[3] `https://xgboost.readthedocs.io/en/stable/`

		Gest. HT/PE	**Gest. DB**	**Total Samples**
History of DB	Train	12.43% / 230	52.97% / 980	– / 1850
	Validation	10.00% / 60	48.33% / 290	– / 600
	Test	12.90% / 80	54.84% / 340	– / 620
History of HT	Train	61.21% / 4450	6.74% / 490	– / 7270
	Validation	61.54% / 1680	5.13% / 140	– / 2730
	Test	61.90% / 1300	7.62% / 160	– / 2100
History of DB+HT	Train	57.30% / 510	29.21% / 260	– / 890
	Validation	62.16% / 230	24.32% / 90	– / 370
	Test	45.95% / 170	45.95% / 170	– / 370
No history of DB/HT	Train	11.18% / 7090	8.01% / 5080	– / 63440
	Validation	11.98% / 2490	8.61% / 1790	– / 20790
	Test	11.45% / 2450	7.99% / 1710	– / 21400

Table 4.3: Summary of sample size and class balance in each subgroup.

4.3.1 Conditioning on Prior History

The top features that the model surfaces include many variants of diabetes and hypertension codes, since prior history of these conditions is highly predictive of gestational diabetes and hypertension. However, there exist a nontrivial number of members who have no prior history of these conditions, and they may be affected by a different set of risk factors. To better model this discrepancy, we partition our dataset, conditional on prior history of diabetes and hypertension, and train a separate model on each subset. We generate 4 subgroups:

- Members with history of diabetes (no hypertension)
- Members with history of hypertension (no diabetes)
- Members with history of diabetes and hypertension
- Members without history of diabetes or hypertension

We report a summary of each subgroup in Table 4.3 and evaluation metrics on the test set in Table 4.4. The subgroup model outperforms the global in accuracy for all subgroups.

		$g^*_{\mathcal{H}}$	g^*
History of DB	AUROC	0.6745	0.7057
	Accuracy	0.6217	0.5700
History of HT	AUROC	0.6573	0.7076
	Accuracy	0.7077	0.6469
History of DB+HT	AUROC	0.6350	0.7568
	Accuracy	0.6243	0.5676
No history of DB/HT	AUROC	0.5948	0.6674
	Accuracy	0.7933	0.7802

Table 4.4: Evaluation metrics for each subgroup, comparing the subgroup model ($g^*_{\mathcal{H}}$) to the global model (g^*).

4.4 Evaluation

4.4.1 Performance Over Time

To assess model performance at different stages of pregnancy, we evaluate the model on subsets of $\mathcal{D}_{test}$, partitioned by where the member is in their pregnancy episode,

$$\mathcal{D}_{test}^{before} = \{(x_t^i, y_t^i) \in \mathcal{D}_{test} | t \geq t_{start}^i - \delta_{buffer} \wedge t < t_{start}^i\} \tag{4.1}$$

$$\mathcal{D}_{test}^{tri1} = \{(x_t^i, y_t^i) \in \mathcal{D}_{test} | t \geq t_{start}^i \wedge t < t_{start}^i + \delta_{tri1}\} \tag{4.2}$$

$$\mathcal{D}_{test}^{tri2} = \{(x_t^i, y_t^i) \in \mathcal{D}_{test} | t \geq t_{start}^i + \delta_{tri1} \wedge t < t_{start}^i + \delta_{tri1} + \delta_{tri2}\} \tag{4.3}$$

$$\mathcal{D}_{test}^{tri3} = \{(x_t^i, y_t^i) \in \mathcal{D}_{test} | t \geq t_{start}^i + \delta_{tri1} + \delta_{tri2} \wedge t < t_{end}^i\} \tag{4.4}$$

where δ_{buffer} is 3 months, δ_{tri1} is 13 weeks, and δ_{tri2} is 14 weeks. Note that we set t_{start}^i to be $t_{start}^{i'} - \gamma$, where $\gamma = 58$ days (mean code delay for start of pregnancy in pregnancies without complications).

We report evaluation metrics on $\mathcal{D}_{test}^{before}$, $\mathcal{D}_{test}^{tri1}$, $\mathcal{D}_{test}^{tri2}$, and $\mathcal{D}_{test}^{tri3}$ in Table 4.5. While confidence intervals do not overlap consistently across time slices, the metrics generally increase as we progress to later pregnancy terms, indicating that the model performs better as we see more data on the member. There is a slight dip in the AUROC from trimester 2 to trimester 3, which may be because the final trimester is more skewed towards members whose pregnancy complications onset later in the

	Accuracy	**AUROC**
Before gestation	0.7309 (0.7216, 0.7402)	0.7219 (0.7116, 0.7322)
Trimester 1	0.7488 (0.7396, 0.7580)	0.7537 (0.7438, 0.7635)
Trimester 2	0.7748 (0.7660, 0.7836)	0.7760 (0.7663, 0.7856)
Trimester 3	0.8159 (0.8077, 0.8239)	0.7687 (0.7573, 0.7801)

Table 4.5: Evaluation metrics for g^*, hybrid predictor for pregnancy risk factors, at different stages of pregnancy with 90% confidence intervals.

pregnancy, and these members are more challenging to predict.

4.4.2 Timeliness of Risk Predictions

To evaluate how early the model is catching pregnancy risk, we plot the distribution of the earliest risk predictions for members at risk of gestational diabetes or hypertension. While the LASSO model has a high false negative rate, of the members with true positive predictions, a majority are caught prior to gestation. This is important since early intervention and treatment are important in reducing gestational diabetes and hypertension risk [55, 46, 48, 26, 58]. We report additional figures in A-2.

4.4.3 Fairness Audit

Prior work has shown that care management risk algorithms may contain racial bias due to nuances in how outcomes are defined [43]. Moreover, there exist systemic health disparities in maternal and infant mortality rates, e.g. Black people have mortality rates over three times higher than White people during pregnancy (40.8 v. 12.7 per 100,000 live births) [44]. To this end, we audit our algorithm for potential racial bias.

We report evaluation metrics in Table 4.6 for the three most common race groups (White - 43.8%, Black - 5.7%, Other[4] - 3.6%). Confidence intervals for AUROC are computed using a distribution-independent method based on error rate and the number of positive and negative samples [16]. Confidence intervals for accuracy are

[4] "Other" race category includes race outside of the following: American Indian or Alaska Native, Black or African American, White, Asian, Hispanic or Latino, Native Hawaiian or Other Pacific Islander.

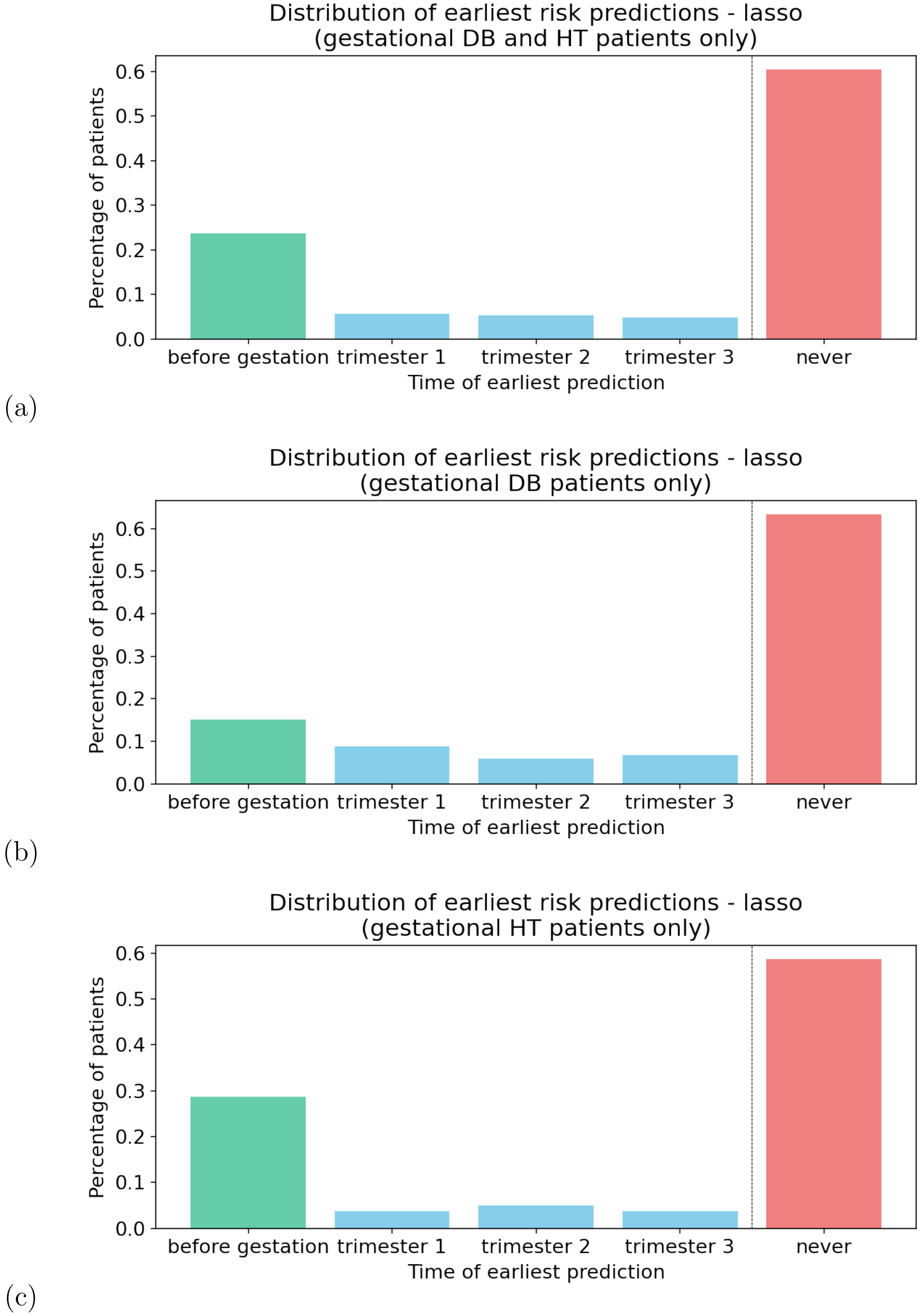


Figure 4-1: Distribution of earliest risk predictions for members at risk of (a) both gestational DB and HT, (b) only gestational DB, and (c) only gestational HT.

computed assuming a Gaussian distribution.

Confidence intervals for error intersect for White and Other race, which indicates comparable performance regardless of race. However, Black members have lower error on average. This may be due to differences in class distribution, since the Black subgroup has much higher rates of complication (44.0%), compared to White (24.6%) and Other (25.9%) race. True positive rates of catching complications are 36.6%, 27.1%, and 30.0%, for Black, White, and Other subgroups, respectively.

Though, we note that data comes from electronic medical records, thus there is low coverage for race attribution (only $\sim 53\%$ of members have some member-level race attributed to examine bias), so we may be misrepresenting error rates across groups.

	Accuracy (90% CI)	**AUROC (90% CI)**
White	0.7745 (0.7680, 0.7811)	0.7401 (0.7321, 0.7481)
Black	0.6806 (0.6596, 0.7015)	0.7866 (0.7683, 0.8049)
Other	0.7918 (0.7689, 0.8147)	0.8262 (0.8026, 0.8498)

Table 4.6: Evaluation metrics for g^*, across different race groups. Rates of complication in each race group are: White - 24.6%, Black - 44.0%, Other - 25.9%.

Chapter 5

Explaining Model Predictions

In this chapter, we motivate requirements for explaining model predictions in the clinical setting. We present several baselines and their limitations, and discuss our proposed explainability method. Finally, we compare explanations for several example members. In a later chapter, we detail a study that examines how explanations for the pregnancy risk model are used by the nurses when integrated into the care management pipeline.

5.1 Motivation

Given our risk model g^* and predicted risk $\hat{y}$, a practitioner may be interested in understanding what features are contributing to, e.g. the gestational diabetes prediction given prior information they know about the member from, e.g. an initial glance at charts. We wish to surface a shortlist of features that we can attribute the prediction to. Due to the redundant nature of healthcare data, however, we wish to avoid surfacing features with redundant information. For example, it is more useful to surface a code for maternal obesity and polycystic ovary syndrome, both risk factors for gestational diabetes, than codes for maternal obesity and morbid obesity.

5.1.1 Requirements

We pose the following requirements for our method:

1. Account for member's **prior knowledge** formalized as a subset of features $X_p \subset X$ that the practitioner already knows about the member, e.g. history of chronic conditions. Our method should not resurface these known covariates.

2. Account for **redundant features** in the data. In healthcare data, some features are *collinear*, i.e. highly correlated with one another. Suppose X^1, X^2 are indicators of lung cancer and cancer, respectively. These features are collinear since $X^1 = 1 \implies X^2 = 1$. In such a case, our method should surface only one of these features.

5.2 Baselines

5.2.1 Global Weights

Given L1-regularized risk model $g^* : R^d \rightarrow \{0, 1, 2\}$, let the probability distribution of class c be parametrized as follows: $p(x = c) = \text{softmax}(\beta_c^\top x)$, where β_c is the parameter for outcome c.

Let d^c be the number of non-zero entries in the weights β_c. We define $\beta_c^{01} : R^d \rightarrow R^{d_c}$ to be a transformation matrix that only keeps the elements of x^i where the corresponding element of β_c is non-zero. Global weights method ranks features according to the magnitude of weights $h(x^i) = \beta_c^{01} x^i$. In other words, we rank features in the support of the member's feature vector by the corresponding weight β_c, for predicted outcome c.

5.2.2 Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a method that approximates a black box predictor with a local linear model to explain each prediction [47]. LIME operates as follows[1]:

[1] https://lime-ml.readthedocs.io/en/latest/lime.html#module-lime.lime_tabular

1. Given an input x^i, generate a set similar points $NE_{x^i}^K$ of size K obtained by perturbing each feature of x^i independently according to the distribution of the training set.

2. Label the set $NE_{x^i}^K$ according to our predictor g^* and learn a sparse linear model h_{x^i} that best performs on the labeled set of similar points. The model is learned using importance weighting based on the similarity between x^i and each point in the neighborhood according to an exponential smoothing kernel.

3. For the predicted class c, rank features according to β_c^i, where the probability distribution of class c for predictor h_{x^i} is: $p(y = c) = \text{softmax}(\beta_c^{i\top} x^i)$.

5.3 Methodology

We introduce two extensions of the baselines in the previous section with modifications to meet the requirements of 5.1.1.

5.3.1 Global Weights++

To satisfy the prior knowledge requirement, we rank features with the model trained on data matching the member's prior history of diabetes and/or hypertension. That is, if member i has prior history of diabetes, we rank features according to $g^*_{\mathcal{H}=DB}$. The redundancy requirement is satisfied by the L1 regularization, since this learns sparse feature weights.

5.3.2 LIME++

To satisfy the prior knowledge requirement, we sample points $NE_{x^i}^K$ according to the distribution of $\mathcal{D}_{train}^{\mathcal{H}=h}$, the subset of training data with prior history h, where member i has prior history of condition h. The redundancy requirement is satisfied by the L1 regularization of the surrogate model, which learns sparse feature weights.

5.4 Results

We show several members for comparison in Figures 5-1-5-5. We observe that global weights and global weights++ give explanations that make sense factually most consistently. While some LIME explanations picked up on useful features, the method is very sensitive to hyperparameter choice such as e.g. kernel width; it may be necessary to use a metric that evaluates explanations to select the optimal hyperparameter. Therefore, we use global weights++ in the user study evaluating pregnancy risk factors in Section 6.2

Feature name	Weight
4047564 - condition - Routine antenatal care (10k day)	-0.4593
4024659 - condition - Gestational diabetes mellitus (10k day)	0.1975
38004461 - specialty - Obstetrics/Gynecology (10k day)	0.1889
2212991 - labs - Culture, presumptive, pathogenic organisms, screening only (10k day)	-0.1720
2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low (365 day)	-0.1615

(a) Global weights – Gestational DB from a previous pregnancy, and specialty visit are surfaced.

Feature name	Weight
4024659 - condition - Gestational diabetes mellitus (10k day)	0.8968
4016041 - condition - Diabetic on diet only (10k day)	0.2433
2212363 - labs - Glucose; tolerance test, each additional beyond 3 specimens (List separately in addition to code for primary procedure) (10k day)	0.2186
2212361 - labs - Glucose; post glucose dose (includes glucose) (30 day)	0.1684
2514533 - procedure - Preventive medicine counseling and/or risk factor reduction intervention(s) provided to an individual (separate procedure); approximately 60 minutes (10k day)	-0.1401

(b) LIME – Gestational DB and associated lab codes from previous pregnancy are surfaced.

Feature name	Weight
4307820 - condition - Unplanned pregnancy (10k day)	0.2146
2414398 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o (10k day)	-0.2021
40762511 - labs - Human papilloma virus 16+18+31+33+35+39+45+51+52+56+58+59+66+68 DNA [Presence] in Cervix by Probe with signal amplification (10k day)	0.1348
2108115 - procedure - Collection of venous blood by venipuncture (10k day)	-0.1261
4024659 - condition - Gestational diabetes mellitus (10k day)	0.1000

(c) Global weights++ – Unplanned pregnancy is associated with increased risk of maternal problems, i.e. it is a proxy for high risk[1]. This, as well as gestational DB from previous pregnancy are surfaced.

Feature name	Weight
2008340 - procedure - Prophylactic administration of vaccine against other diseases (10k day)	0.0076
3013157 - labs - Ampicillin+Sulbactam [Susceptibility] (10k day)	0.0055
765719 - condition - Lump in lower outer quadrant of left breast (365 day)	0.0040
3023143 - labs - Ciprofloxacin [Susceptibility] (10k day)	0.0036
3004202 - labs - Nitrofurantoin [Susceptibility] (10k day)	0.0028

(d) LIME++ – Various antibiotics and treatments for bacterial infections are surfaced, but are not predictive of gestational DB.

Figure 5-1: Member predicted to be gestational diabetic, with history of gestational DB and pre-eclampsia from a previous pregnancy.

Feature name	Weight
4047564 - condition - Routine antenatal care (730 day)	-0.4593
438480 - condition - Abnormal glucose tolerance in mother complicating pregnancy, childbirth AND/OR puerperium (730 day)	0.2386
4024659 - condition - Gestational diabetes mellitus (365 day)	0.1975
38004461 - specialty - Obstetrics/Gynecology (365 day)	0.1889
2212991 - labs - Culture, presumptive, pathogenic organisms, screening only (365 day)	-0.1720

(a) Global weights – Gestational DB, abnormal glucose, and specialty visit from a previous pregnancy are surfaced.

Feature name	Weight
4024659 - condition - Gestational diabetes mellitus (365 day)	0.6970
438480 - condition - Abnormal glucose tolerance in mother complicating pregnancy, childbirth AND/OR puerperium (730 day)	0.6481
320128 - condition - Essential hypertension (10k day)	-0.2497
2212363 - labs - Glucose; tolerance test, each additional beyond 3 specimens (List separately in addition to code for primary procedure) (10k day)	0.2304
4016041 - condition - Diabetic on diet only (365 day)	0.2151

(b) LIME – Gestational DB and associated lab codes from previous pregnancy are surfaced, as well as HT (as a negative factor).

Feature name	Weight
2414398 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o (730 day)	-0.2021
2108115 - procedure - Collection of venous blood by venipuncture (365 day)	-0.1261
0 - race - No matching concept (nontemporal)	-0.1232
4024659 - condition - Gestational diabetes mellitus (365 day)	0.1077
4016041 - condition - Diabetic on diet only (365 day)	0.1000

(c) Global weights++ – Gestational DB codes from a previous pregnancy are surfaced.

Feature name	Weight
437623 - condition - Polyhydramnios (365 day)	0.0059
4014716 - condition - Placental finding (365 day)	0.0058
2212652 - labs - Blood count; red blood cell (RBC), automated (10k day)	0.0048
4143187 - condition - Anomaly of placenta (365 day)	-0.0024
4062557 - condition - False labor (365 day)	0.0023

(d) LIME++ – Polyhydraminos, excess of amniotic fluid, which is highly associated with diabetes, is surfaced. [25, 40]

Figure 5-2: Member predicted to be gestational diabetic, with history of gestational DB from a previous pregnancy and chronic hypertension.

Feature name	Weight
2212545 - labs - Protein, total, except by refractometry; urine (180 day)	0.1777
42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum (180 day)	0.0694
2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low (180 day)	-0.0491
379805 - condition - Myopia (10k day)	-0.0457
2212648 - labs - Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (180 day)	-0.0446

(a) Global weights – HT lab code and maternal obesity (a risk factor) are surfaced.

Feature name	Weight
2212545 - labs - Protein, total, except by refractometry; urine (180 day)	0.5366
42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum (180 day)	0.1695
2212295 - labs - Creatinine; other source (180 day)	0.0713
3024561 - labs - Albumin [Mass/volume] in Serum or Plasma (180 day)	0.0489
38004461 - specialty - Obstetrics/Gynecology (30 day)	-0.0306

(b) LIME – HT lab codes and maternal obesity (a risk factor) are surfaced.

Feature name	Weight
2212545 - labs - Protein, total, except by refractometry; urine (180 day)	0.1038
42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum (180 day)	0.0829
3020876 - labs - Protein [Mass/time] in 24 hour Urine (180 day)	0.0745
2212295 - labs - Creatinine; other source (180 day)	0.0575
2414398 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o (365 day)	-0.0500

(c) Global weights++ – HT lab codes and maternal obesity (a risk factor) are surfaced.

Feature name	Weight
40162447 - drug - 0.5 ML ganirelix acetate 0.5 MG/ML Prefilled Syringe (180 day)	0.0137
1113346 - drug - aspirin 81 MG Chewable Tablet (180 day)	0.0089
19127938 - drug - 7 (ethinyl estradiol 0.025 MG / norgestimate 0.18 MG Oral Tablet) / 7 (ethinyl estradiol 0.025 MG / norgestimate 0.215 MG Oral Tablet) / 7 (ethinyl estradiol 0.025 MG / norgestimate 0.25 MG Oral Tablet) / 7 (inert ingredients 1 MG Oral Tablet) Pack [Or (10k day)	0.0084
3039154 - labs - Gestational age Estimated from conception date (30 day)	0.0020
46235242 - labs - Fetal Trisomy 18 risk [Interpretation] based on Plasma cell-free+WBC DNA by Dosage of chromosome-specific cfDNA Qualitative (30 day)	0.0012

(d) LIME++ – Fertility treatment (ganirelix) and birth control (ethinyl estradiol) are surfaced, but are not predictive of gestational HT.

Figure 5-3: Member predicted to be gestational hypertensive, with pre-existing hypertension. No history of prior pregnancy and has been on fertility treatments.

Feature name	Weight
2212991 - labs - Culture, presumptive, pathogenic organisms, screening only (180 day)	-0.1706
8516 - race - Black or African American (nontemporal)	0.1017
137940 - condition - Transient hypertension of pregnancy - delivered (180 day)	0.0975
439393 - condition - Pre-eclampsia (180 day)	0.0899
3001582 - labs - Protein/Creatinine [Mass Ratio] in Urine (180 day)	0.0836

(a) Global weights – Black race, associated with HT for environmental and behavioral factors [20], is surfaced, as well as history of HT and HT labs.

Feature name	Weight
3001582 - labs - Protein/Creatinine [Mass Ratio] in Urine (180 day)	0.3667
439393 - condition - Pre-eclampsia (180 day)	0.3332
8516 - race - Black or African American (nontemporal)	0.2211
137940 - condition - Transient hypertension of pregnancy - delivered (180 day)	0.2086
2101814 - procedure - Anesthesia for cesarean delivery following neuraxial labor analgesia/anesthesia (List separately in addition to code for primary procedure performed) (180 day)	-0.1254

(b) LIME – Black race, associated with HT for environmental and behavioral factors, is surfaced, as well as history of HT and HT labs.

Feature name	Weight
3019897 - labs - Erythrocyte distribution width [Ratio] by Automated count (180 day)	-0.1147
2101814 - procedure - Anesthesia for cesarean delivery following neuraxial labor analgesia/anesthesia (List separately in addition to code for primary procedure performed) (180 day)	-0.1065
8516 - race - Black or African American (nontemporal)	0.0758
3001582 - labs - Protein/Creatinine [Mass Ratio] in Urine (180 day)	0.0428
4047564 - condition - Routine antenatal care (180 day)	-0.0331

(c) Global weights++ – Black race, associated with HT for environmental and behavioral factors, is surfaced, as well as HT labs.

Feature name	Weight
3034426 - labs - Prothrombin time (PT) (180 day)	0.0061
3022250 - labs - Lactate dehydrogenase [Enzymatic activity/volume] in Serum or Plasma by Lactate to pyruvate reaction (180 day)	-0.0043
3035511 - labs - Protein [Mass/volume] in Urine collected for unspecified duration (180 day)	0.0033
2101813 - procedure - Neuraxial labor analgesia/anesthesia for planned vaginal delivery (this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor) (180 day)	0.0031
3016407 - labs - Fibrinogen [Mass/volume] in Platelet poor plasma by Coagulation assay (180 day)	0.0021

(d) LIME++ – Various HT tests surfaced, though lactate dehydrogenase is surfaced as a negative predictor.

Figure 5-4: Member predicted to be gestational hypertensive, with pre-existing hypertension, advanced maternal age.

Feature name	Weight
4047564 - condition - Routine antenatal care (365 day)	0.2136
Age at end date (nontemporal)	0.1546
8527 - race - White (nontemporal)	0.0465
2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low (365 day)	0.0331
2212648 - labs - Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (180 day)	0.0257

(a) Global weights – Routine care, visit, and bloodwork are surfaced. White race, which is associated with lower pregnancy risk [44], is also surfaced.

Feature name	Weight
4047564 - condition - Routine antenatal care (365 day)	0.1844
2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low (365 day)	0.0691
38004461 - specialty - Obstetrics/Gynecology (180 day)	-0.0552
38004450 - specialty - Anesthesiology (365 day)	0.0414
2514527 - procedure - Periodic comprehensive preventive medicine reevaluation and management of an individual including an age and gender appropriate history, examination, counseling/anticipatory guidance/risk factor reduction interventions, and the ordering of laboratory/diag (180 day)	0.0268

(b) LIME – Routine care and visit are surfaced. Specialty OB visit is surfaced as a negative predictor.

Feature name	Weight
2212996 - labs - Culture, bacterial; quantitative colony count, urine (180 day)	0.3662
8527 - race - White (nontemporal)	0.1587
4047564 - condition - Routine antenatal care (365 day)	0.1454
2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low (365 day)	0.1130
Age at end date (nontemporal)	0.0672

(c) Global weights++ – Routine care, visit, and labwork are surfaced. White race, which is associated with lower pregnancy risk, is also surfaced.

Feature name	Weight
135287 - condition - Non-neoplastic nevus (730 day)	0.0080
136057 - condition - Benign neoplasm of skin of trunk (730 day)	0.0057
2110326 - procedure - Treatment of missed abortion, completed surgically; first trimester (365 day)	0.0047
2211459 - procedure - Radiologic examination, hand; minimum of 3 views (365 day)	0.0035
4141481 - condition - Enteroviral vesicular stomatitis with exanthem (730 day)	0.0024

(d) LIME++ – Conditions that are not associated with pregnancy risk are surfaced, though not very informative with respect to pregnancy. Missed abortion treatment is also surfaced.

Figure 5-5: Member predicted to have no complication, with no prior history of DB or HT.

Chapter 6

User Study

Deploying machine learning-based clinical decision support tools is challenging, and it is important to study how well these systems work when working collaboratively with human clinicians. In this chapter, we outline two user studies aimed at evaluating how (1) the early identification of pregnancy algorithm, and (2) pregnancy risk factor model and explanations compare to the existing clinical workflow. The following studies are exempted from Institutional Review Board (IRB) approval; we detail the exemption category in A.5.

6.1 Early Identification of Pregnancy

The objective of the study is to evaluate whether the hybrid pregnancy identification algorithm can help nurses identify pregnant members earlier than the codes alone, without introducing previously unseen error (e.g. on non-pregnant members). To this end, we sample members from three categories: members detected pregnant early by the model, members detected pregnant by the codes, and members who are never pregnant. For the model-detected and non-pregnant categories, we also include false positive / false negative members to evaluate whether the nurse can correctly filter members incorrectly identified by the algorithm.

We also evaluate whether an explanation of the model's prediction aids nurses' decision making. Therefore, we run two trials – one with model's pregnancy inference and explanations, and one without.

6.1.1 Study Setup

Sampling members. For each trial, we sample 12 members distributed as shown in Table 6.1. We sample pregnant members from the pregnancies without complications sub-cohort, since this gives a less noisy label for pregnancy start. We manually checked pregnant members who were detected too early by the model, to ensure it was not due to, e.g. positive pregnancy signal from a previous pregnancy.

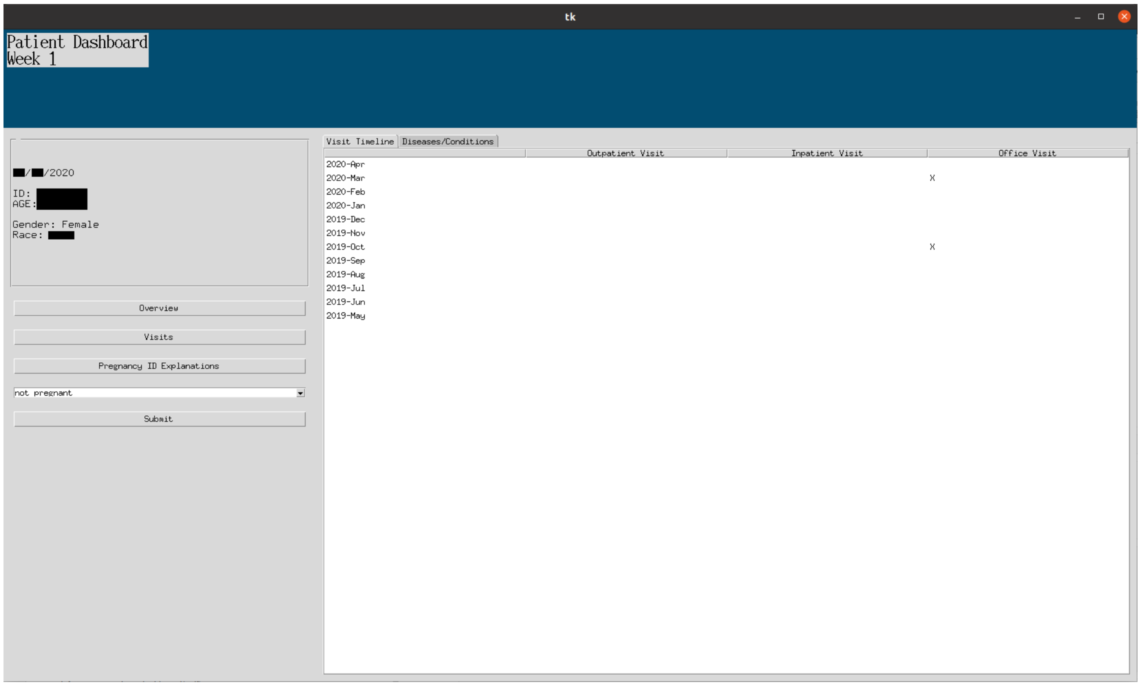
Simulating data. We simulate data and algorithm output for these members for up to 5 weeks, or until the nurse labels the member pregnant. To ensure the model or code triggers for members predicted pregnant within the simulation window, we sample the simulation start date around the inferred pregnancy start date. We report the sampling window in Table 6.1. The pregnancy identification algorithm is run on data up to 4 weeks back, for the present week and previous week. Note that we simulate a 30-day claims lag by omitting any data within 30 days of the present date. We generate model explanations using global weights method.

Member Dashboard UI. To run the study, we built a mock member dashboard to surface medical history available in insurance claims and other data sources (e.g. visits, diagnosis codes, demographics) for nurses to learn about how the new model and explainability tools impact performance such as speed and accuracy of member identification and outreach. Examples of the interface are shown in Figures 6-1-6-4.

Collecting data. Each week, for each member (who has not yet been labeled pregnant), we asked the nurse if they think the member is pregnant or not pregnant. We measured the total time taken to parse through the members' data to answer this question.

Category	**Sub-category**	**# of members**	**Simulation start date range**
Pregnant members detected by model	Detected early within reasonable time (at least 1 month after t_{start})	2	$[t^*_{start} - 3 \text{ weeks}, t^*_{start}]$
	Detected too early (before 1 month after t_{start})	2	$[t^*_{start} - 3 \text{ weeks}, t^*_{start}]$
Pregnant members detected by code	–	4	$[t^*_{start} - 1 \text{ week}, t^*_{start} + 2 \text{ weeks}]$
Non-pregnant members	Detected not pregnant	3	$\left[\tau^0, \tau^{'} - 5 \text{ weeks}\right]$
	Detected pregnant	1	$[t^*_{start} - 3 \text{ weeks}, t^*_{start}]$

Table 6.1: Distribution of members and range of simulation start dates for pregnancy identification study. Note that t^*_{start} is the pregnancy start time inferred by the algorithm, and τ^0 and $\tau^{'}$ are the first and last time points in data sampled for a member.



Patient Dashboard
Week 1

[redacted]/[redacted]/2020

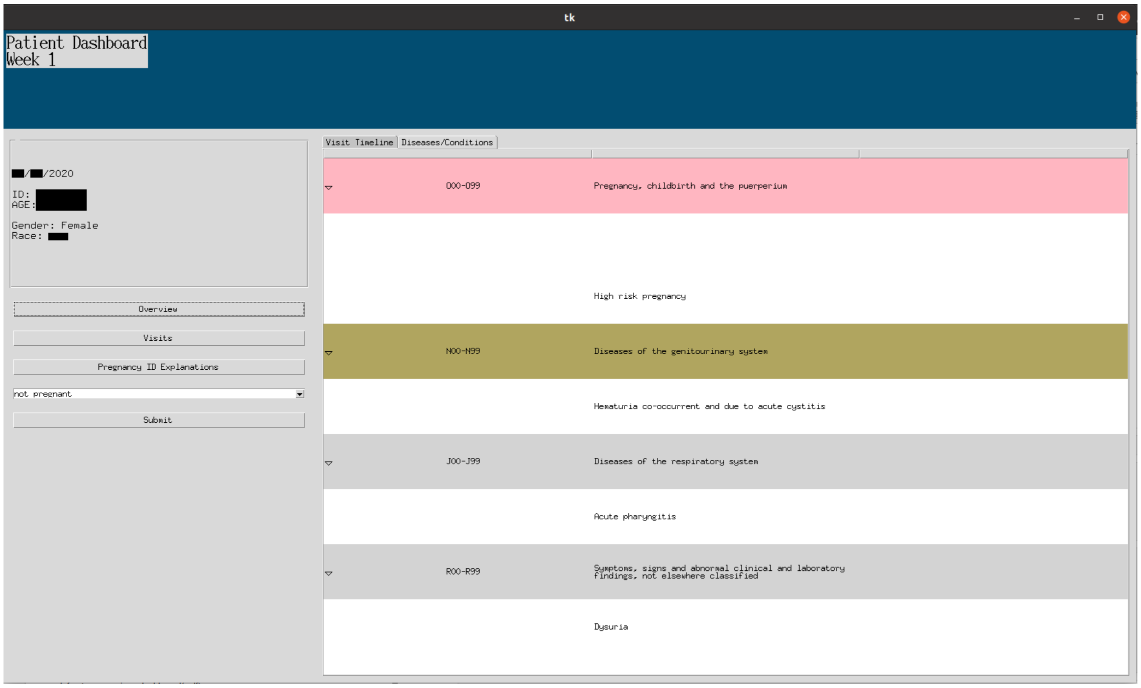
ID:
AGE: [redacted]

Gender: Female
Race: [redacted]

not pregnant

	Outpatient Visit	Inpatient Visit	Office Visit
2020-Apr			
2020-Mar			X
2020-Feb			
2020-Jan			
2019-Dec			
2019-Nov			
2019-Oct			X
2019-Sep			
2019-Aug			
2019-Jul			
2019-Jun			
2019-May			

Figure 6-1: Visit Timeline – Summary of visit types for the past year of visits.



Patient Dashboard
Week 1

[redacted]/[redacted]/2020

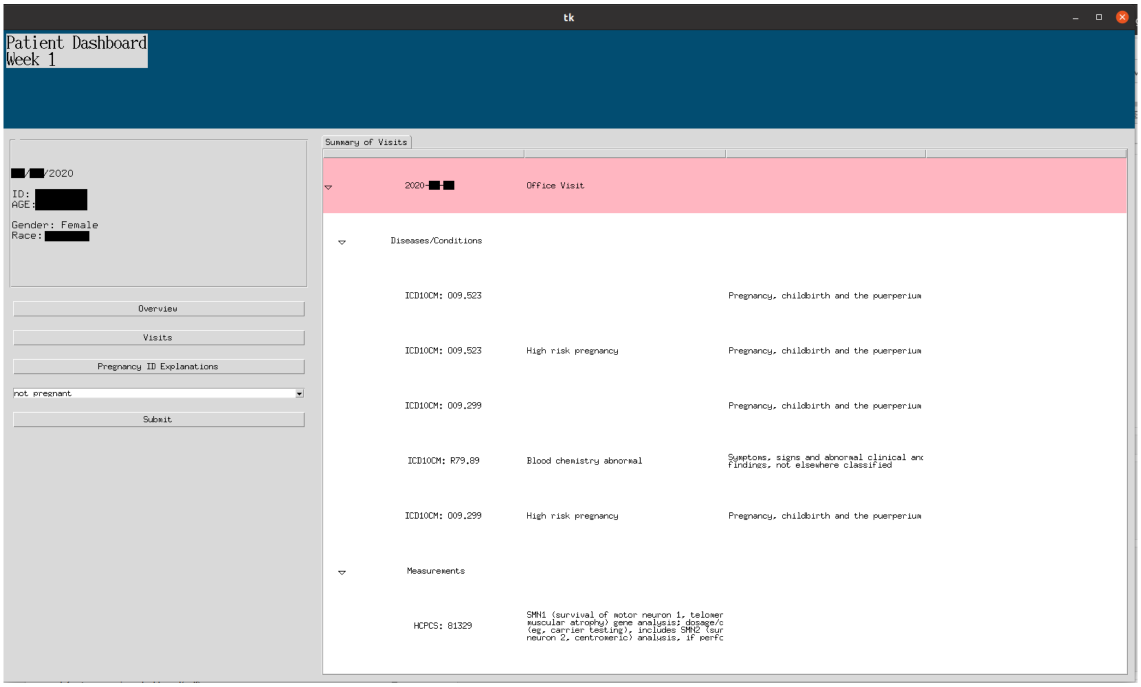
ID:
AGE: [redacted]

Gender: Female
Race: [redacted]

not pregnant

▽	O00-O99	Pregnancy, childbirth and the puerperium
		High risk pregnancy
▽	N00-N99	Diseases of the genitourinary system
		Hematuria co-occurrent and due to acute cystitis
▽	J00-J99	Diseases of the respiratory system
		Acute pharyngitis
▽	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
		Dysuria

Figure 6-2: Diseases/Conditions – Summary of diagnosis codes, categorized by ICD 10 code ranges.

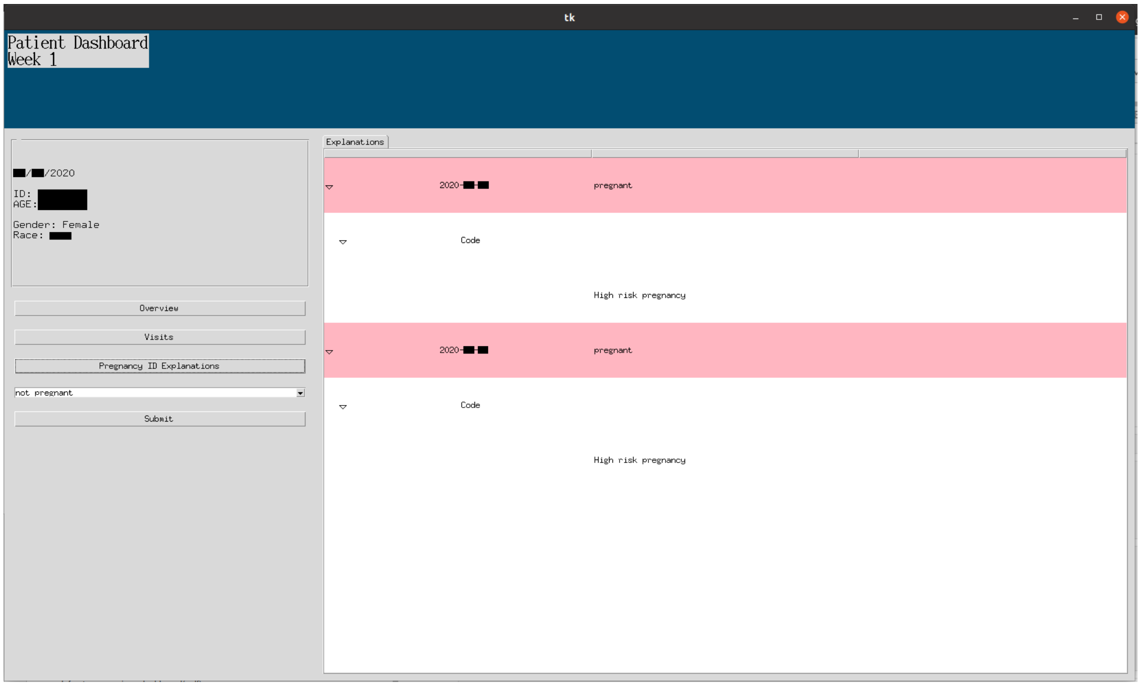

tk

Patient Dashboard
Week 1

[redacted]/[redacted]/2020
ID: [redacted]
AGE:
Gender: Female
Race: [redacted]

2020-[redacted]-[redacted]	Office Visit	
Diseases/Conditions		
ICD10CM: O09.523		Pregnancy, childbirth and the puerperium
ICD10CM: O09.523	High risk pregnancy	Pregnancy, childbirth and the puerperium
ICD10CM: O09.299		Pregnancy, childbirth and the puerperium
ICD10CM: R79.89	Blood chemistry abnormal	Symptoms, signs and abnormal clinical anc findings, not elsewhere classified
ICD10CM: O09.299	High risk pregnancy	Pregnancy, childbirth and the puerperium
Measurements		
HCPCS: 81329	SMN1 (survival of motor neuron 1, telomer muscular atrophy) gene analysis; dosage/c (eg, carrier testing), includes SMN2 (sur neuron 2, centromeric) analysis, if perfc	

Figure 6-3: Summary of Visits – Member's clinical codes for each visit, categorized by type of data.


tk

Patient Dashboard
Week 1

[redacted]/[redacted]/2020
ID: [redacted]
AGE:
Gender: Female
Race: [redacted]

2020-[redacted]-[redacted]	pregnant
Code	
	High risk pregnancy
2020-[redacted]-[redacted]	pregnant
Code	
	High risk pregnancy

Figure 6-4: Explanations – Pregnancy inference for past two weeks, as well as code(s) surfaced by the algorithm.

6.1.2 Results

We report a summary of results – (1) number of members labeled pregnant and (2) how many days earlier or later pregnancy is detected by nurses, relative to the codes – per category in tables 6.2 and 6.3 for **trial A** (without prediction or explanation) and **trial B** (with prediction and explanation). We also report a distribution of the time taken to label members for each trial in Table 6.4. Since both trials label everyone correctly, except 1 (pregnant) person, classification metrics are identical. We report these metrics in Table 6.5.

6.1.3 Discussion

Nurses labeled pregnant members with good sensitivity and specificity, as both trials only mislabeled 1 (pregnant) person. However, we observed that the explanations caused the nurse to overthink, when she disagreed with the model outcome, or when the explanation wasn't exactly indicative of pregnancy, even if there were other indications of pregnancy in the full visit summary. This is supported in Table 6.4, which shows that nurses spent longer time sorting through pregnant members when presented with model explanations (trial B).

In trial A, we observe that the average days for the nurse to label pregnant members, relative to the code is earlier for members surfaced as pregnant by the model, than for members surfaced by the code. We also observe that all false negative and false positive members were properly filtered out. This provides evidence for the model's deploy-ability since the model can surface pregnant members earlier, which the nurse correctly validates, while filtering out members surfaced incorrectly. However, trial B suggests that the explanation method and *how* we display explanations must be improved for pregnancy identification.

One suggestion for displaying explanations was to better integrate them into the clinical workflow by highlighting surfaced features on the "Visits" page, which displays clinical codes for all visits. Often times, a single code alone isn't sufficient to conclude pregnancy, but viewing related codes during the same visit can give more context. We

Category	**Sub-category**	**# of members (total)**	**# of members labeled pregnant**	**# days pregnancy is detected, relative to pregnancy start code**
Pregnant members detected by model	Detected early within reasonable time (at least 1 month after t_{start})	2	1	avg. 74 days early
	Detected too early (before 1 month after t_{start})	2	0	–
Pregnant members detected by code	–	4	4	avg. 5 days late
Non-pregnant members	Detected not pregnant	3	0	–
	Detected pregnant	1	0	–

Table 6.2: Summary of results for trial A (pregnancy identification without prediction or explanation).

also received general feedback on missing features in our dashboard: lab measurements (with abnormal measurements highlighted), indication of specialty visits, and top diagnosis codes for each visit; we took this into account for the next user study.

6.2 Predicting and Explaining Pregnancy Risk Factors

The objective of this study is to evaluate (1) if model predictions can aid enrollment decision-making in pregnancy care management, and (2) if model explanations can help nurses develop more insight into the member before outreach, and see if these behaviors generalize across users. To this end, we run three trials – one without predictions or explanations, one with predictions, and one with both predictions and explanations, with each of the two nurses from the pregnancy care management program (six trials total).

Category	**Sub-category**	**# of members (total)**	**# of members labeled pregnant**	**# days pregnancy is detected, relative to pregnancy start code**
Pregnant members detected by model	Detected early within reasonable time (at least 1 month after t_{start})	2	2	avg. 16 days late
	Detected too early (before 1 month after t_{start})	2	0	–
Pregnant members detected by code	–	4	3	avg. 16 days late
Non-pregnant members	Detected not pregnant	3	0	–
	Detected pregnant	1	0	–

Table 6.3: Summary of results for trial B (pregnancy identification with prediction and explanation).

Trial		μ	σ
A	all members	36.1 s	16.5 s
	pregnant only	45.1 s	12.9 s
	non-pregnant only	31.7 s	16.2 s
B	all members	37.8 s	24.2 s
	pregnant only	54.8 s	28.2 s
	non-pregnant only	26.5 s	11.1 s

Table 6.4: Distribution of time to label members for each trial.

Trial	**Acc.**	**TPR**	**FPR**	**PPV**	**NPV**
A	0.92	0.83	1.00	1.00	0.86
B	0.92	0.83	1.00	1.00	0.86

Table 6.5: General classification metrics, computed from nurses' pregnancy labels.

Outcome	Correct Prediction?	Prior History?	Number of Members
Gestational DB	Yes	No DB history	3
Gestational HT	Yes	No HT history	3
No complication	Yes	No DB or HT history	3
Gestational DB	No	No DB history	1
Gestational HT	No	No HT history	1
No complication	No	No DB or HT history	1
Gestational DB	Yes	DB history	1
Gestational HT	Yes	HT history	1
No complication	Yes	DB+HT history	1
Gestational DB	No	DB history	1
Gestational HT	No	HT history	1
No complication	No	DB+HT history	1

Table 6.6: Distribution of members for pregnancy risk factor study.

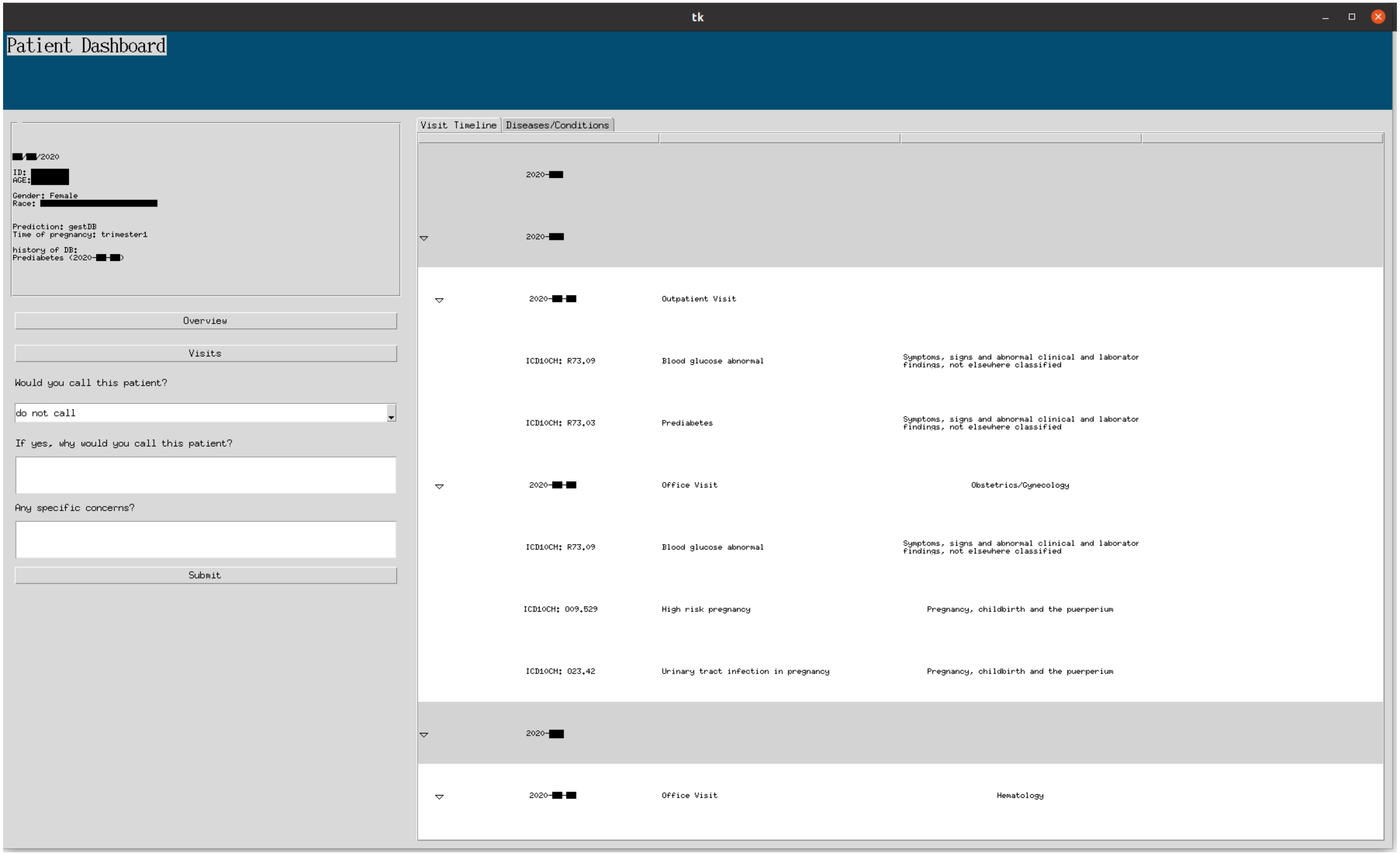
6.2.1 Study Setup

Sampling members. For each trial, we sample 18 members distributed as shown in Table 6.6. We sample members with and without prior history, with more members without prior history to reflect the dataset distribution. We include false positives and negatives to assess nurses' ability to filter incorrect predictions.

Simulating data. For each member, we sample a data point uniformly from 3 months before pregnancy start to pregnancy end. We simulate 30-day claims lag by omitting any data within 30 days of the present date from the dashboard, model input, etc. We obtain model prediction from g^* and generate model explanations using global weights++ method.

Member Dashboard UI. We used a modified version of the mockup of the member dashboard used by the care management nurses from the pregnancy identification study. We added lab measurements to the "Summary of Visits" tab, with indication of abnormal values. We modified the "Visit Timeline" tab to indicate any specialty visits and display diagnosis codes for each visit. Finally, we removed the separate page for model explanations and integrate them into the "Summary of Visits" tab by highlighting codes that are positively (green) or negatively (red) associated with the prediction. Examples of the interface are shown in Figures 6-5-6-7.

Collecting data. For each member, we asked the nurse if they would call the member, and if they are calling, why. We measured the total time taken to parse through the members' data to answer these questions. For the call list question, we gave five options: do not call, call for gestational diabetes, call for gestational hypertension, call for both gestational diabetes and hypertension, and call for another risk factor. We included the last option to account for additional risk factors, but did not count it as a label concordant with gestational diabetes or hypertension labels.



Patient Dashboard

[redacted]/[redacted]/2020

ID: [redacted]
AGE:

Gender: Female
Race: [redacted]

Prediction: gestDB
Time of pregnancy: trimester1

history of DB:
Prediabetes (2020-[redacted]-[redacted])

Would you call this patient?

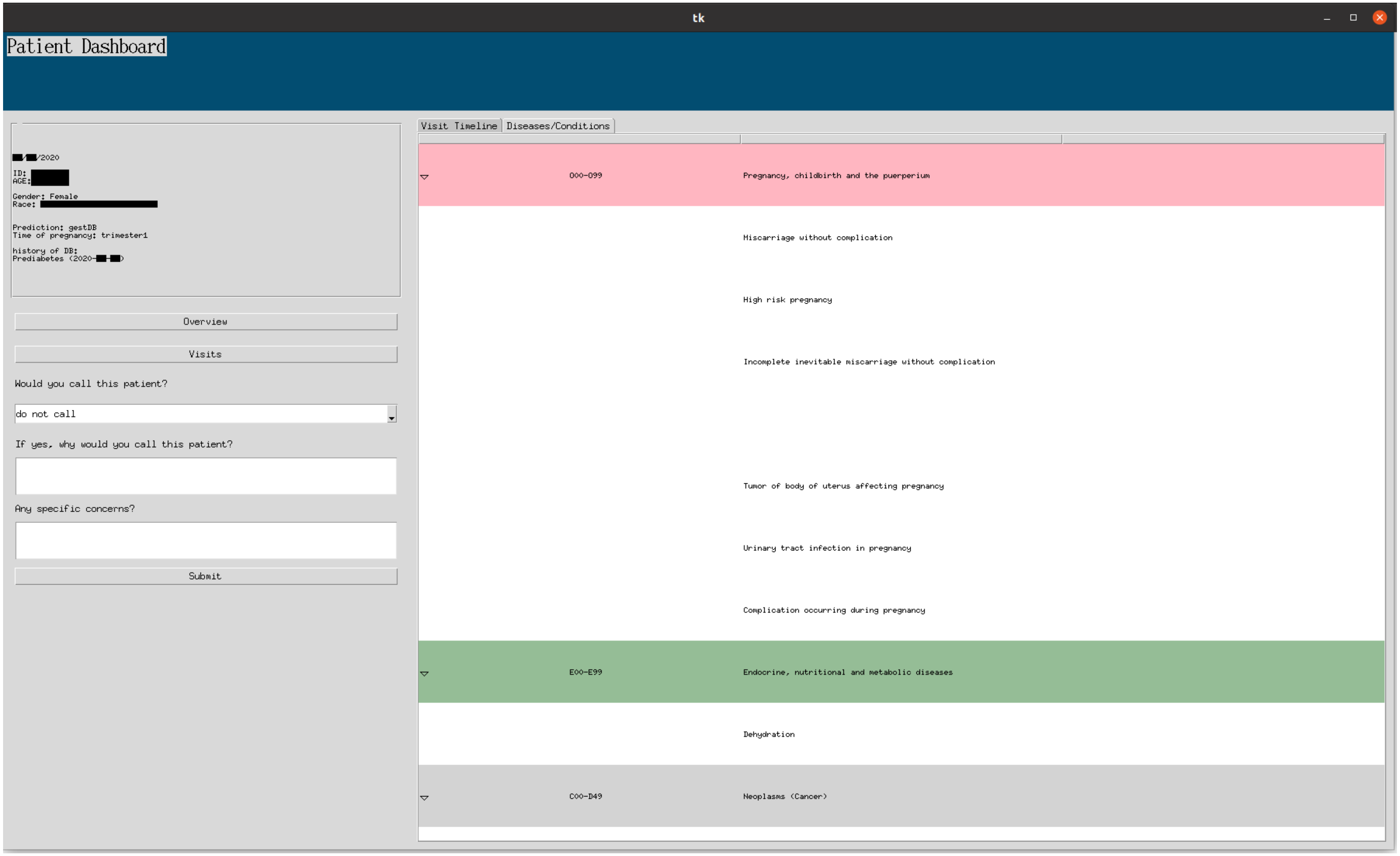
do not call

If yes, why would you call this patient?

Any specific concerns?

2020-[redacted]		
2020-[redacted]		
2020-[redacted]-[redacted]	Outpatient Visit	
ICD10CM: R73.09	Blood glucose abnormal	Symptoms, signs and abnormal clinical and laborator findings, not elsewhere classified
ICD10CM: R73.03	Prediabetes	Symptoms, signs and abnormal clinical and laborator findings, not elsewhere classified
2020-[redacted]-[redacted]	Office Visit	Obstetrics/Gynecology
ICD10CM: R73.09	Blood glucose abnormal	Symptoms, signs and abnormal clinical and laborator findings, not elsewhere classified
ICD10CM: O09.529	High risk pregnancy	Pregnancy, childbirth and the puerperium
ICD10CM: O23.42	Urinary tract infection in pregnancy	Pregnancy, childbirth and the puerperium
2020-[redacted]		
2020-[redacted]-[redacted]	Office Visit	Hematology

Figure 6-5: Visit Timeline (v2) – Summary of visits for the past year. Shows visit type and diagnosis codes for each visit and indicates specialty visits (e.g. OB/GYN, Hematology).



Patient Dashboard

[redacted]/2020

ID: [redacted]
AGE:

Gender: Female
Race: [redacted]

Prediction: gestDB
Time of pregnancy: trimester1

history of DB:
Prediabetes (2020-[redacted])

Would you call this patient?

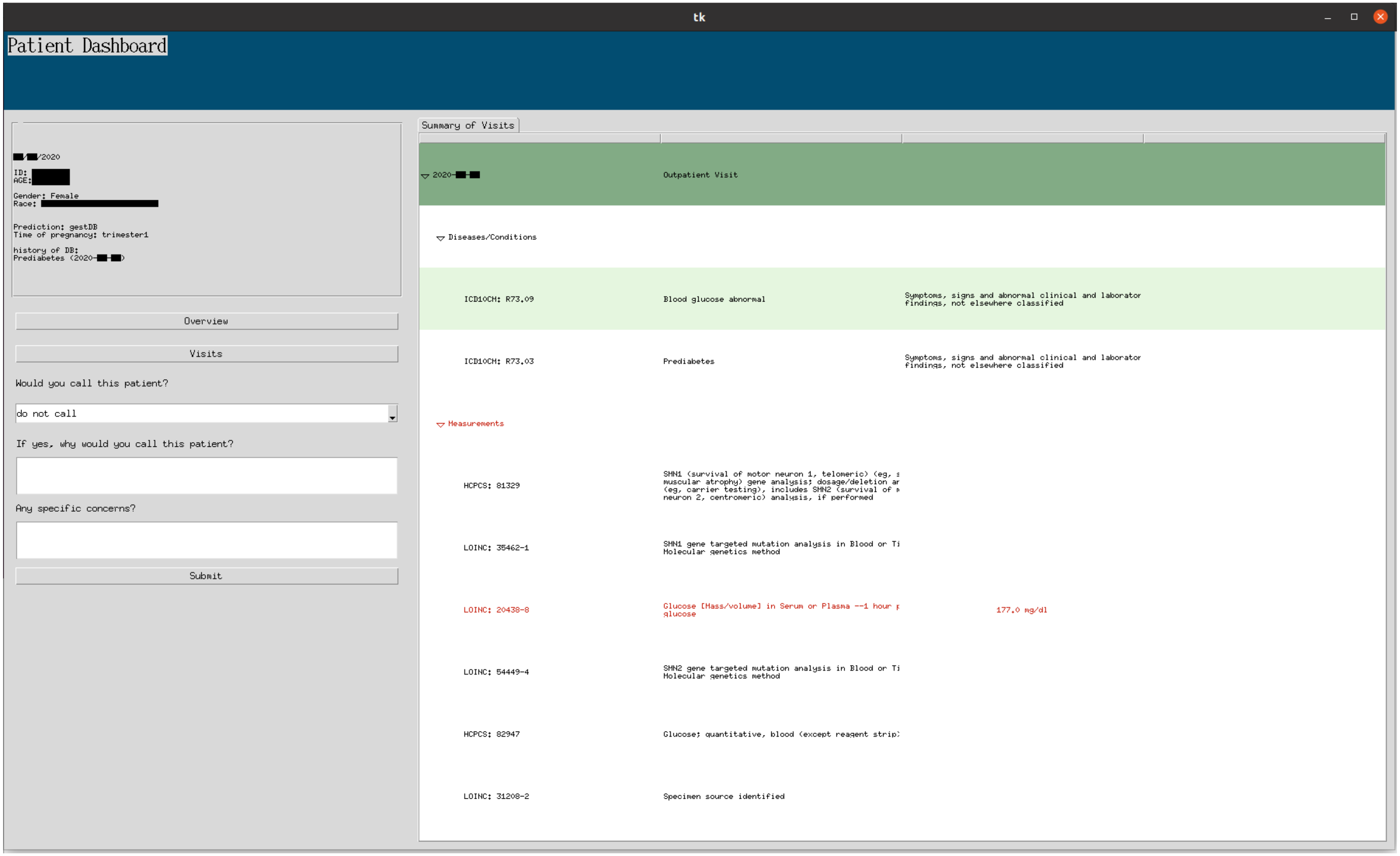
do not call

If yes, why would you call this patient?

Any specific concerns?

▽	O00-O99	Pregnancy, childbirth and the puerperium
		Miscarriage without complication
		High risk pregnancy
		Incomplete inevitable miscarriage without complication
		Tumor of body of uterus affecting pregnancy
		Urinary tract infection in pregnancy
		Complication occurring during pregnancy
▽	E00-E99	Endocrine, nutritional and metabolic diseases
		Dehydration
▽	C00-D49	Neoplasms (Cancer)

Figure 6-6: Diseases/Conditions (v2) – Summary of diagnosis codes, categorized by ICD 10 code ranges.



Patient Dashboard

[redacted]/[redacted]/2020

ID: [redacted]
AGE:

Gender: Female
Race: [redacted]

Prediction: gestDB
Time of pregnancy: trimester1

history of DB:
Prediabetes (2020-[redacted]-[redacted])

Would you call this patient?

do not call

If yes, why would you call this patient?

Any specific concerns?

Summary of Visits

▽ 2020-[redacted]-[redacted]	Outpatient Visit		
▽ Diseases/Conditions			
ICD10CM: R73.09	Blood glucose abnormal	Symptoms, signs and abnormal clinical and laborator findings, not elsewhere classified	
ICD10CM: R73.03	Prediabetes	Symptoms, signs and abnormal clinical and laborator findings, not elsewhere classified	
▽ Measurements			
HCPCS: 81329	SMN1 (survival of motor neuron 1, telomeric) (eg, s muscular atrophy) gene analysis; dosage/deletion ar (eg, carrier testing), includes SMN2 (survival of m neuron 2, centromeric) analysis, if performed		
LOINC: 35462-1	SMN1 gene targeted mutation analysis in Blood or Ti Molecular genetics method		
LOINC: 20438-8	Glucose [Mass/volume] in Serum or Plasma --1 hour p glucose	177.0 mg/dl	
LOINC: 54449-4	SMN2 gene targeted mutation analysis in Blood or Ti Molecular genetics method		
HCPCS: 82947	Glucose; quantitative, blood (except reagent strip)		
LOINC: 31208-2	Specimen source identified		

Figure 6-7: Summary of Visits (v2) – Member's clinical codes for each visit, categorized by type of data. Abnormal lab measurements are in red (high) or blue (low). Codes surfaced by the explanation method are highlighted in green (positive) or red (negative), depending on how they are correlated with the model prediction.

(a)

Trial	Mean	St. Dev.
A	102.4 s	41.4 s
B	86.1 s	30.3 s
C	95.2 s	32.4 s

(b)

Trial	Mean	St. Dev.
A	115.4 s	55.6 s
B	82.7 s	37.9 s
C	109.0 s	43.1 s

Table 6.7: Distribution of time for nurse to decide whether to call member and provide explanation for nurse 1 (a) and nurse 2 (b).

(a)

Trial	Acc. (↑)	TPR (↑)	FPR (↓)	PPV (↑)	NPV (↑)
A	0.56	0.75	0.67	0.69	0.40
B	0.72	0.75	0.33	0.82	0.57
C	0.67	0.83	0.50	0.77	0.60

(b)

Trial	Acc. (↑)	TPR (↑)	FPR (↓)	PPV (↑)	NPV (↑)
A	0.33	0.33	0.50	0.57	0.27
B	0.56	0.67	0.67	0.67	0.33
C	0.67	0.75	0.50	0.75	0.50

Table 6.8: General classification metrics, computed from nurses' call labels for nurse 1 (a) and nurse 2 (b).

6.2.2 Results

We report results for **trial A** (no prediction, prior history, or explanations), **trial B** (prediction and prior history included, no explanations), and **trial C** (prediction, prior history, and explanations included), for each nurse (nurse 1 and nurse 2). We report summary statistics on time taken by each nurse to filter through members in Table 6.7. We also report classification metrics across the three trials, based on the nurses' call labels in Table 6.8, and metrics on agreement between the nurses' call labels and (correct) model predictions in Table 6.9. Finally, we do a quantitative evaluation of the nurses' notes explaining the reason for their calls, by flagging notes that contain information beyond just the prior knowledge (age, race, prior history of diabetes/hypertension). We report these metrics and examples in Table 6.10.

6.2.3 Discussion

Although time to sort through members are comparable across the trials for both nurses, we observe that inclusion of model prediction and prior history (trial A versus

(a)

Trial	Agreement with prediction
A	0.25
B	0.33
C	0.33

(b)

Trial	Agreement with prediction
A	0.58
B	0.25
C	0.25

Table 6.9: Proportion of nurses' call labels that agree with model prediction, for correctly predicted members for nuse 1 (a) and nurse 2 (b).

B) allow nurses to spend less time per member across both nurses (nurse 1: p=0.09, nurse 2: p=0.02), as we report in Table 6.7, though part of the improvement may be due to nurses becoming more familiar with the task as trials progressed. Inclusion of prediction and prior history improve accuracy, TPR-FPR tradeoff, PPV, and NPV, for at least one of the two trials (trial B or C) across both nurses (Table 6.8). We observed that prior history was heavily relied upon in trials B and C.

In Table 6.9, we observe greater agreement between nurses' call labels and (correct) model predictions when model prediction was presented for nurse 1, which may suggest that model predictions can aid some nurses' decision making, though not to a statistically significant degree (trial B: p=0.35, trial C: p=0.35). For nurse 2, we observe the opposite trend, that there is less agreement (trial B: p=0.01, trial C: p=0.01). This suggests that a good deferral algorithm (e.g. defer all members whose predictions are likely incorrect) may further enhance performance metrics reported in Table 6.8, depending on the nurse.

When we include model explanations (trial B versus C), nurses were less reluctant to parse through the history of members' clinical codes and thus included more information beyond prior knowledge (nurse 1: p=0.004, nurse 2: p=0.003), Table 6.10, at the expense of increased time per member (nurse 1: p=0.19, nurse 2: 0.03). Nurse 1 explained that prior history of diabetes/hypertension or complications in previous pregnancy are usually sufficient to make a call, but additional information such as distinct risk factors for complications (e.g. polycystic ovary syndrome, cer-

(a)

Trial	Members with information beyond prior knowledge	Examples
A	55.6% (10 members)	oligohydramnios, premature delivery, blood clot, cervical issues, large baby, fetal hereditary disease, cancer, abnormal heart rate, first pregnancy
B	33.3% (6 members)	mental health / potential for postpartum depression, fetal abnormality, obesity, cervical issue, first pregnancy
C	66.7% (12 members)	elevated glucose during current pregnancy / abnormal glucose code, h/o premature delivery, first pregnancy, twins, polycystic ovary syndrome, cervical incompetence (risk for preterm birth), elevated protein labs in current pregnancy

(b)

Trial	Members with information beyond prior knowledge	Examples
A	55.6% (10 members)	previous retained placenta, home injections, pulmonary embolism, pre-term delivery, thalassemia, elevated glucose, asthma, hypothyroidism, infertility, uterine leimyoma, anemia, musculoskeletal disease, polycystic ovary syndrome, methadone
B	27.8% (5 members)	asthma, pre-term delivery, hypothyroidism, obesity, fibroids, infertility
C	61.1% (11 members)	firbroids, previous losses, pre-term delivery, thyroid disease, cardiac murmur, Rhesus -, obesity, cardiac concern, hypothyroidism, infertility

Table 6.10: Summary of members whose nurse notes contain information beyond prior knowledge (age, race, prior history of DB/HT) for nurse 1 (a) and nurse 2 (b).

vical incompetence) or elevated labs from recent visits can help them build a better profile of the member and identify those at immediate risk. Both nurses indicated that prior history and highlighted explanations helped with obtaining this information more quickly. The explanations helped them focus in on important visits and codes, especially when the visit history was lengthy. Nurse 2 said that although not all explanations were useful or made sense, it is easy to filter out the unuseful ones, i.e. surfacing useful codes should be prioritized over surfacing few codes.

Chapter 7

Discussion

In this thesis, we have introduced new components for identifying pregnant members, predicting members at risk of developing pregnancy risk factors, and explaining these models' predictions, for the high risk pregnancy care management enrollment pipeline at Independence Blue Cross. We set up a mock enrollment dashboard and evaluated these methods across two user studies and found that (1) the pregnancy identification algorithm helps nurses identify pregnancies earlier while correctly filtering out false positive members, and (2) showing the model's prediction and prior history of chronic conditions improves nurse's performance metrics when deciding who to call. While model explanations adversely affected the nurse's performance in terms of time per member and how early they identify pregnant members in the pregnancy identification study, we observed that explanations improved notes about the member in the pregnancy risk factor study without much difference in nurse's classification performance. The latter study better integrated explanations into the clinical workflow, and nurses appeared to disagree with the explanations less, which emphasizes the importance of the explanation method and how they are presented.

Though we made an initial attempt at outlining requirements for an explainability framework in clinical settings, our methods lack a causal angle that is necessary for a more nuanced framework. When explaining clinical models, we want the explanations to inform interventions that can reverse the outcome, which suggests a treatment effect-esque approach of measuring the importance of each feature. We elaborate on this in Section 7.1.1.

7.1 Future Works

7.1.1 Explaining Model Predictions in Clinical Settings

In this work, we focused mostly on the the requirements of prior knowledge about the member and accounting for redundancy in data. Given the healthcare application, another avenue to explore is to add a causal angle to the method. To motivate, consider the following – given our risk model g^* and predicted risk $\hat{y}$, a nurse may be interested in understanding:

1. Why should this member be enrolled in the care management program?, and
2. What factors are leading to outcome $\hat{y}$ that the member can reduce their impact?

While an answer to the first question might include something like high age, the second question asks specifically about factors that can be intervened on to reduce or reverse the downstream outcome – that is, it would not include features like age or history of pregnancy complications since they cannot be intervened upon. Instead, the second question may point to factors like the member's high blood pressure measurements; the nurse can then provide education to monitor their blood pressure closely and adjust their diet accordingly to reverse their downstream outcome of developing e.g. gestational hypertension.

With this causal angle, we can also expand our redundancy requirement to consider the dependency structure of redundant features so we surface root causes over proxies for those variables. For example, if we have indicators of ovarian cancer and cancer, we would like our method to surface ovarian cancer over cancer since this is more informative in understanding the member and understanding treatment options.

Conditional Average Treatment Effect (CATE)

One way to formalize this notion is by measuring the "treatment effect" of intervening on each feature and surfacing the features with the largest effect on the outcome. Let

Y^c represent the risk outcome c, $\mathbf{X^P}$ represent the subset of prior knowledge features, and X^f represent the feature of interest. Then the conditional average treatment effect (CATE) of intervening on feature f for a person with feature x^f and prior information $\mathbf{x^P}$ is:

$$CATE(x^f, \mathbf{x^P}) = E\left[Y^c \mid \mathrm{do}(X^f) = x^f, \mathbf{X^P} = \mathbf{x^P}\right] - E\left[Y^c \mid \mathrm{do}(X^f) = \bar{x^f}, \mathbf{X^P} = \mathbf{x^P}\right]. \tag{7.1}$$

Given a given member with features $\mathbf{x}$, we have features that are present, $\mathbf{x_1}$, and features that are absent, $\mathbf{x_0}$, since all features are binary. It is usually sufficient for the purpose of the nurses to find the presence of a few features that indicate that the member is high-risk. This is because the absence of a feature may be because it wasn't recorded in the data (the member didn't have a specific test or procedure for one reason or the other). Thus, we only compute CATE for the present features $\mathbf{x_1}$.

Does CATE address our requirements?

1. **Prior knowledge:** By conditioning on $\mathbf{X^P}$, CATE should address the requirement of prior knowledge. However, conditioning on prior knowledge may not always affect the treatment effect estimate, depending on the parametrization of the outcome. We expand on this issue in "Effect of conditioning on prior knowledge".

2. **Redundant features:** Since CATE is computed independently for each feature, if two features are identical then they will have the same CATE value. Thus they will have the same ranking, and this needs to be remedied. We talk about a potential fix in section "Subsetwise CATE". However, if feature a and feature b have a particular dependence structure, e.g. feature a causes feature b, then we want to surface feature a over feature b, but this may not necessarily follow, as we will describe in "CATE and Underlying Causes".

Effect of conditioning on prior knowledge. We illustrate an example to show that depending on the parametrization of the outcome, the treatment effect may or may not be affected by the prior knowledge we condition on.

Suppose we have X^1, X^2 are indicators of lung cancer and cancer, respectively, and X^3 is an indicator of high age (greater than 40). We assume that X^3 and the pair X^1, X^2 are not connected in the structural causal model, i.e. high age does not cause cancer, and clearly cancer doesn't cause high age.

1. **Case 1:** Suppose the outcome is linear: $Y = \beta^1 X^1 + \beta^2 X^2 + \beta^3 X^3$. Then, the average treatment effect (ATE) and CATE evaluate to the same value:

$$\begin{aligned}
ATE(1) &= E[Y|\mathrm{do}(X^1) = 1] - E[Y|\mathrm{do}(X^1) = 0] \\
&= (\beta^1 + \beta^2 + \beta^3 P(X^3 = 1)) - (\beta^2 P(X^2 = 1|\mathrm{do}(X^1) = 0) + \beta^3 P(X^3 = 1)) \\
&= \beta^1 + \beta^2(1 - P(X^2 = 1|\mathrm{do}(X^1) = 0)) \\
CATE(1) &= E[Y|\mathrm{do}(X^1) = 1, X_3 = x_3] - E[Y|\mathrm{do}(X^1) = 0, X_3 = x_3] \\
&= (\beta^1 + \beta^2 + \beta^3 x^3) - (\beta^2 P(X^2 = 1|\mathrm{do}(X^1) = 0) + \beta^3 x^3) \\
&= \beta^1 + \beta^2(1 - P(X^2 = 1|\mathrm{do}(X^1) = 0))
\end{aligned}$$

   In other words, conditioning on the member age does not impact any estimate we may compute for CATE.

2. **Case 2:** Suppose the outcome is non-linear: $Y = \beta^1 X^1 + \beta^2 X^2 + \beta^3 X^3 + \beta^{1,3} X^1 X^3 - \delta$, where the non-linearity comes from the interaction term $X^1 X^3$. Note that $\delta > 0$, and we interpret $Y > 0$ to mean a member has outcome c, and $Y < 0$ to mean that the member does not have the outcome. The magnitude of Y indicates severity of the outcome. The average treatment effect evaluates

to the following:

$$\begin{aligned}
ATE(1) &= E[Y|\mathrm{do}(X^1) = 1] - E[Y|\mathrm{do}(X^1) = 0] \\
&= (\beta^1 + \beta^2 + \beta^3 Pr(X^3 = 1) + \beta^{1,3} Pr(X^3 = 1|\mathrm{do}(X^1) = 1)) \\
&- (0 + \beta^2 Pr(X^2 = 1|\mathrm{do}(X^1) = 0) + \beta^3 Pr(X^3 = 1) + \beta^{1,3} Pr(X^3 = 1|\mathrm{do}(X^1) = 0)) \\
&= (\beta^1 + \beta^2 + \beta^3 Pr(X^3 = 1) + \beta^{1,3} Pr(X^3 = 1)) \\
&- (0 + \beta^2 Pr(X^2 = 1|\mathrm{do}(X^1) = 0) + \beta^3 Pr(X^3 = 1) + \beta^{1,3} Pr(X^3 = 1)) \\
&= \beta^1 + \beta^2 (1 - P(X^2 = 1|\mathrm{do}(X^1) = 0))
\end{aligned}$$

If we condition on high age, i.e. $X^3 = 1$, our treatment effect evaluates to the following:

$$\begin{aligned}
CATE(1) &= E[Y|\mathrm{do}(X^1) = 1, X^3 = 1] - E[Y|\mathrm{do}(X^1) = 0, X^3 = 1] \\
&= (\beta^1 + \beta^2 + \beta^3 + \beta^{1,3}) - (0 + \beta^2 Pr(X^2 = 1|\mathrm{do}(X^1) = 0) + \beta^3 + 0) \\
&= \beta^1 + \beta^2 (1 - P(X^2 = 1|\mathrm{do}(X^1) = 0)) + \beta^{1,3}
\end{aligned}$$

Thus, our estimate for the effect of intervening on lung cancer is now higher, which we may change the feature rankings.

A key caveat here, is which subset of features to condition to get the feature importance estimate. Here, one might say X^3 makes sense since it is independent from X^1 in the structural causal model. Generally speaking, we should not condition on any post-treatment variables (e.g. X^2) to get our estimate since that will bias our results.

CATE and Underlying Causes. We present a few examples to illustrate that CATE may not always surface root underlying causes.

Example 1. We use the following motivating example, where X^1, X^2 are indicators of lung cancer and cancer, respectively. We want to predict risk outcome Y. We assume we have an unobserved variable X^3 denoting breast cancer and that $X^2 = X^1 \vee X^3$: cancer node is 1 if we have breast cancer or lung cancer. Assume $Y = \beta^1 X^1 + \beta^2 X^2$.

Interpret β^2 as the effect of cancer that is not special to lung cancer and interpret β^1 as the effect of lung cancer that is not capture by other cancers. Here it makes sense possibly to have $\beta^1 < \beta^2$, assuming regularization. The average treatment effect (ATE) of intervening on cancer is as follows:

$$\begin{aligned}
ATE(2) &= E[Y|\text{do}(X^2) = 1] - E[Y|\text{do}(X^2) = 0] \\
&= E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^2) = 1] - E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^2) = 0] \\
&= \beta^2 + \beta^1 E[X^1|do(X^2) = 1] - \beta^2 \cdot 0 - \beta^1 \cdot E[X^1|do(X^2) = 0] \\
&= \beta^2 + \beta^1 \cdot (E[X^1|do(X^2) = 1] - E[X^1|do(X^2) = 0]) \\
&= \beta^2
\end{aligned}$$

The last line holds as X^2 is downstream of X^1 and so the distribution of X^1 is unaffected by the operation $do(X^2) = 1$. The ATE of intervening of lung cancer is:

$$\begin{aligned}
ATE(1) &= E[Y|\text{do}(X^1) = 1] - E[Y|\text{do}(X^1) = 0] \\
&= E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^1) = 1] - E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^1) = 0] \\
&= \beta^2 + \beta^1 - \beta^1 \cdot 0 - \beta^2 \cdot [X^2|do(X^1) = 0] \\
&= \beta^1 + \beta^2(1 - P(X^3 = 1))
\end{aligned}$$

For the given member if $X^3 = 0$, then clearly the lung cancer has a higher ATE, if $X^3 = 1$ then the effect of intervening on lung cancer can be lower than the effect of cancer. Here cancer is the proxy node while lung cancer is the root cause.

Example 2. Suppose $X_1 \to X_2 \to Y$, and all variables are binary. We have $P(X_2 = 1|X_1) = c \cdot X_1$ and $P(Y = 1|X_2) = b \cdot X_2$. Then the the ATE of X_2 is c. However, the ATE of the "root cause", X_1, is $c \cdot b$ which is smaller than c because everything is in $[0, 1]$.

Example 3. Let X^1 denote hypertension and let X^2 be kidney disease indicator and X^3 be a drug used to treat hypertension, where the causal graph is: $X_1 \to X_2, X_1 \to X_3$.

We want to predict risk outcome $Y = \beta_1 X_1 + \beta_2 X_2$. The ATE of kidney disease is:

$$\begin{aligned}
ATE(2) &= E[Y|\text{do}(X^2) = 1] - E[Y|\text{do}(X^2) = 0] \\
&= E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^2) = 1] - E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^2) = 0] \\
&= \beta^2 + \beta^1 [X^1|do(X^2) = 1] - \beta^2 \cdot 0 - \beta^1 \cdot [X^1|do(X^2) = 0] \\
&= \beta^2 + \beta^1 \cdot ([X^1|do(X^2) = 1] - [X^1|do(X^2) = 0]) \\
&= \beta^2
\end{aligned}$$

The last line holds as X^2 is upstream from X^1 and so the distribution of X^1 is unaffected by the operation $do(X^2) = 1$. The ATE of hypertension is:

$$\begin{aligned}
ATE(1) &= E[Y|\text{do}(X^1) = 1] - E[Y|\text{do}(X^1) = 0] \\
&= E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X^1) = 1] - E[\beta^1 X^1 + \beta^2 X^2|\text{do}(X1) = 0] \\
&= \beta^1 + \beta^2 P(X_2 = 1|do(X_1) = 1) - \beta^1 \cdot 0 - \beta^2 \cdot E[X^2|do(X^1) = 0] \\
&= \beta^1 + \beta^2 (P(X_2 = 1|do(X_1) = 1) - P(X_2 = 1|do(X_1) = 0))
\end{aligned}$$

The ATE of the hypertension drug is:

$$\begin{aligned}
ATE(3) &= E[Y|\text{do}(X^3) = 1] - E[Y|\text{do}(X^3) = 0] \\
&= 0
\end{aligned}$$

Subsetwise CATE. The main issue with using the CATE values and ranking the features is that two identical features will have the same rank and value. Suppose we only want to select K features to display. One way is to look for the subset $S \subset X$ that has the highest treatment effect:

$$S^*(x^P) = \tag{7.2}$$

$$\arg\max_{S \subset X_1, |S|=K} E\left[Y^c \mid \text{do}(S) = 1, X^P = x^P\right] - E\left[Y^c \mid \text{do}(S) = 0, X^P = x^P\right], \tag{7.3}$$

where x^P is the set of prior knowledge features for the member of interest. To include member-specific features, consider the following modification:

$$S^*(x^P) = \tag{7.4}$$

$$\arg\max_{S \subset X_1, |S|=K} \quad E\left[Y \mid \text{do}(S) = 1, X^P = x^P, X_1 \setminus S = x_1 \setminus s\right] - \tag{7.5}$$

$$E\left[Y \mid \text{do}(S) = 0, X^P = x^P, X_1 \setminus S = x_1 \setminus s\right], \tag{7.6}$$

where $X_1 \setminus S$ represents the covariates beyond the set S and $do(S) = 1$ means setting all the variables in S are all equal to 1.

Since this is a joint optimization over the features, we will only surface non-redundant features. Furthermore, this subset formulation allows us to measure, e.g. pairwise effects of variables, unlike CATE which is computed independently for each feature.

Appendix A

Appendix

A.1 Cohort Creation

The modeling problems defined in Chapters 3 and 4 rely on data from pregnant and non-pregnant patients. Central to constructing these datasets are cohort creation algorithms that identifies pregnant and non-pregnant patients, which we detail in Sections A.1.1 and A.1.2, respectively.

A.1.1 Building Pregnant Cohort

We build on [38], which presents an algorithm for inferring pregnancy episodes across a set of pregnancy outcomes in OMOP Common Data Model. Our modified algorithm can handle a larger set of pregnancy outcomes, e.g. neonatal ICU admission, by doing a forward search to update the outcome once the pregnancy episode is identified. We describe our modified version in Algorithm 3. We illustrate the algorithm in Figure 3-1 and present a subset of target codes for reference in A.2.

A.1.2 Building Non-Pregnant Cohort

We build a cohort of patients who were never pregnant throughout their claims history. We sample these patients according to the age distribution of pregnant members (mean: 31.8 years, standard deviation: 4.8 years) and define "never pregnant" to be

Algorithm 3 Building pregnant cohort.

```
for $i \in P$ do
    // Detect and classify most recent pregnancy outcome (first pass)
    $t^i_{out}, outcome^i \leftarrow getPregnancyOutcome(\mathcal{H}^i)$

    // Backtrack to estimate pregnancy start
    $t_{start,min} \leftarrow t^i_{out} - g^{outcome}_{max}$          ▷ Lower bound for pregnancy start
    $t_{start,max} \leftarrow t^i_{out} - g^{outcome}_{min}$          ▷ Upper bound for pregnancy start
    $t^i_{start} \leftarrow estimatePregnancyStart(t_{start,min}, t_{start,max})$

    // Forward search to update pregnancy outcome (second pass)
    $t^i_{out}, outcome^i \leftarrow updatePregnancyOutcome(t^i_{start}, t^i_{out})$
end for
```

any member who does not have any of the pregnancy start or outcome concept codes present in their claims history.

A.2 Target Definitions

In Algorithm 3, the first pass phase that searches for the most recent pregnancy outcome references the original pregnancy outcomes and corresponding target codes defined in [38]. In the second pass phase that performs a second search to update the previous outcome, we reference target codes for additional outcomes. We present a subset of target codes for these outcomes and indicator for when they are used in Table A.1

Outcome	**Target Codes**	**Pregnancy ID?**	**Risk Factors?**
Neonatal Intensive Care Unit (NICU)	Newborn light for gestational age	X	X
	Low birth weight infant		
	Birth injury to central nervous system		
	Respiratory distress syndrome in the newborn		
	Pulmonary hypertension of newborn		
Hypertension/Pre-eclampsia (HPPE)	Pre-existing hypertension in obstetric context	X	X
	Transient hypertension of pregnancy		
	Renal hypertension complicating pregnancy		
	Severe pre-eclampsia		
	Gestational proteinuria		
Pre-term birth	Preterm premature rupture of membranes	X	X
	Fetal or neonatal effect of maternal premature rupture of membrane		
	Baby premature, 24-26 weeks		
	Extreme immaturity, 750-999 grams		
	Metabolic bone disease of prematurity		
Gestational Hypertension	Unspecified maternal hypertension		X
	Gestational [pregnancy-induced] hypertension		
	Hypertension, Pregnancy-Induced		
	gestational proteinuria		
	Mild to moderate pre-eclampsia		
Gestational Diabetes	Gestational diabetes mellitus in childbirth		X
	Diabetes mellitus arising in pregnancy		
	Gestational diabetes mellitus in the puerperium		
	Gestational diabetes mellitus complicating pregnancy		
	Maternal gestational diabetes mellitus		

Table A.1: Pregnancy outcomes and examples of corresponding target codes and indicators of whether the outcome was included in the second pass search during cohort creation for pregnancy identification and pregnancy risk factors.

Hyperparameters	**Search Range**
Regularization strength (C)	1e-3, 7.5e-4, 5e-4, 2.5e-4*, 1e-4
Tolerance	1*, 1e-1, 1e-2, 1e-3, 1e-4

Table A.2: Hyperparameter search range for pregnancy identification model. Asterisk marks the chosen hyperparameters.

	Hyperparameters	**Search Range**
LASSO	Regularization strength (C)	1, 1e-1, 1e-2, 1e-3*, 1e-4
	Tolerance	1e-1, 1e-2, 1e-3*, 1e-4
ELASTIC-NET	L1-ratio	0.25*, 0.5, 0.75
	Tolerance	1e-1*, 5e-2
XGBOOST	Learning rate	1e-1*, 1e-2, 1e-3, 1e-4

Table A.3: Hyperparameter search range for pregnancy risk model. Asterisk marks the chosen hyperparameters.

A.3 Training Configurations

A.3.1 Pregnancy Identification Model

We report the hyperparameter search space in Table A.2. We select the model with the highest validation accuracy. The decision threshold is chosen to be the geometric mean of sensitivity and specificity on the validation set.

A.3.2 Pregnancy Risk Factor Model

We report the hyperparameter search space in Table A.3. Note that we also correct for class imbalance by weighting each class j by $p(y_j)^{-1}$, where $p(y_j)$ is the proportion of outcomes under class j in the training set. We select the model with highest product of AUROC and accuracy on the validation set.

A.4 Evaluations

A.4.1 Pregnancy Identification

Model features

We report the top 25 weighted features in the pregnancy identification classifier, f^* in Table A.4.

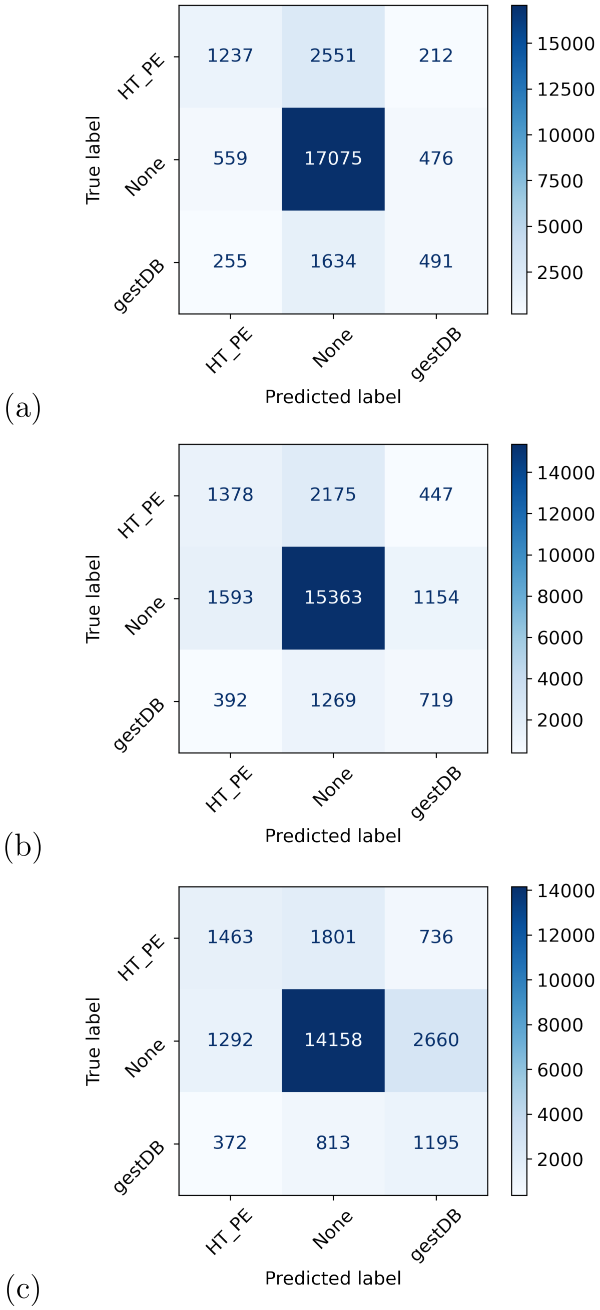
Table A.4: Top 25 features of pregnancy identification classifier, f^*, by weight.

Feature Type	# of Features	Top Features (+)	Top Features (-)
Demographics	12	–	–
Conditions	19,266	45765728 - condition - Supervision of high risk pregnancy, 0.2398 4047564 - condition - Routine antenatal care, 0.1509 43530950 - condition - Complication occurring during pregnancy, 0.0727 43530881 - condition - Suspected fetal disorder, 0.0520 4111608 - condition - Normal fetal growth, 0.0368 42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum, 0.0339 72693 - condition - Poor fetal growth affecting management	
Procedure	18,226	2108115 - procedure - Collection of venous blood by venipuncture, 0.0869 2213418 - procedure - Immunization administration (includes percutaneous, intradermal, subcutaneous, or intramuscular injections); 1 vaccine (single or combination vaccine/toxoid), 0.0713	2110307 - procedure - Routine obstetric care including antepartum care, vaginal delivery (with or without episiotomy, and/or forceps) and postpartum care, -0.0362
Labs	13,912	2212361 - labs - Glucose; post glucose dose (includes glucose), 0.1673 40757116 - labs - Culture, typing; identification by nucleic acid (DNA or RNA) probe, amplified probe technique, per culture or isolate, each organism probed, 0.0812 2212167 - labs - Urinalysis, by dip stick or tablet reagent for bilirubin, glucose, hemoglobin, ketones, leukocytes, nitrite, pH, protein, specific gravity, urobilinogen, any number of these constituents; non-automated, without microscopy, 0.0801 3025315 - labs - Body weight, 0.0783 2212802 - labs - Inhibin A, 0.0691 3048882 - labs - Streptococcus agalactiae DNA [Presence] in Unspecified specimen by NAA with probe detection, 0.0670 2212996 - labs - Culture, bacterial; quantitative colony count, urine, 0.0662 2212991 - labs - Culture, presumptive, pathogenic organisms, screening only, 0.0513 3011564 - labs - Rubella virus IgG Ab [Units/volume] in Serum or Plasma by Immunoassay, 0.0453 3037110 - labs - Fasting glucose [Mass/volume] in Serum or Plasma, 0.0451 3050479 - labs - Immature granulocytes/100 leukocytes in Blood, 0.0343	3006923 - labs - Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma, -0.0518
Drugs	11,100	42800265 - drug - 0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML / Bordetella pertussis fimbriae 2/3 vaccine, inactivated 0.01 MG/ML / Bordetella pertussis pertactin vaccine, inactivated 0.006 MG/ML / Bordetella pertussis toxoid vacci, 0.0508 43526402 - drug - doxylamine succinate 10 MG / pyridoxine hydrochloride 10 MG Delayed Release Oral Tablet [Diclegis], 0.0359	–
Specialty	218	–	38004456 - specialty - Internal Medicine, -0.0467

A.4.2 Pregnancy Risk Factor

Confusion Matrices

We report confusion matrices for the global risk factor models, for each classification algorithm in Figure A-1.



(a)

(b)

(c)

Figure A-1: Confusion matrices for pregnancy risk model, (a) LASSO, (b) ELASTIC-NET, (c) XGBOOST.

Model features

We report the top 25 weighted features for each outcome (gestational diabetes, gestational hypertension, and no complication) in the LASSO (L1-regularized) pregnancy risk factor model, g^* in Tables A.5-A.7.

We also report the top 25 weighted features for each subgroup model conditioned on prior history in Tables A.8-A.10 ($g^*_{\mathcal{H}=DB}$), A.11-A.13 ($g^*_{\mathcal{H}=HT}$), A.14-A.16 ($g^*_{\mathcal{H}=DB+HT}$), and A.17-A.19 ($g^*_{\mathcal{H}=none}$).

Table A.5: Top 25 features of LASSO pregnancy risk factor model, g^*, ranked by gestational diabetes weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12	8515 - race - Asian, 0.1383	Age at end date, -0.0936
Conditions	37,795	438480 - condition - Abnormal glucose tolerance in mother complicating pregnancy, childbirth AND/OR puerperium - 10000 days, 0.2415 433736 - condition - Obesity - 10000 days, 0.1979 4024659 - condition - Gestational diabetes mellitus - 10000 days, 0.1975 4132434 - condition - Gestation period, 8 weeks - 180 days, 0.1056 4320944 - condition - Cellulitis of toe - 10000 days, 0.0871 4042728 - condition - Blood glucose abnormal - 730 days, 0.0860 138113 - condition - Cyst of thyroid - 10000 days, 0.0854	4047564 - condition - Routine antenatal care - 180 days, -0.1921 320128 - condition - Essential hypertension - 10000 days, -0.1312 36713926 - condition - Somatic dysfunction of thoracic region - 10000 days, -0.1274 4047564 - condition - Routine antenatal care - 365 days, -0.0991 4047564 - condition - Routine antenatal care - 730 days, -0.0980 138525 - condition - Pain in limb - 365 days, -0.0860
Procedure	27,635		2110307 - procedure - Routine obstetric care including antepartum care, vaginal delivery (with or without episiotomy, and/or forceps) and postpartum care - 10000 days, -0.1101 2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low - 730 days, -0.1047 2101813 - procedure - Neuraxial labor analgesia/anesthesia for planned vaginal delivery (this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor) - 730 days, -0.0942 2514527 - procedure - Periodic comprehensive preventive medicine reevaluation and management of an individual including an age and gender appropriate history, examination, counseling/anticipatory guidance/risk factor reduction interventions, and the ordering of laboratory/diag - 730 days, -0.0935
Labs	26,815	2212363 - labs - Glucose; tolerance test, each additional beyond 3 specimens (List separately in addition to code for primary procedure) - 10000 days, 0.1394 2212361 - labs - Glucose; post glucose dose (includes glucose) - 30 days, 0.1057 3019762 - labs - Thyrotropin [Units/volume] in Serum or Plasma by Detection limit <= 0.05 mIU/L - 730 days, 0.1016	2212991 - labs - Culture, presumptive, pathogenic organisms, screening only - 30 days, -0.1336
Drugs	19,550		
Specialty	515	38004461 - specialty - Obstetrics/Gynecology - 10000 days, 0.1889	

Table A.6: Top 25 features of LASSO pregnancy risk factor model, g^*, ranked by gestational hypertension weight.

Feature Type	# of Features	Top Features (+)	Top Features (-)
Demographics	12	8516 - race - Black or African American, 0.1017	8522 - race - Other Race, -0.1601 8515 - race - Asian, -0.1126
Conditions	37,795	320128 - condition - Essential hypertension - 10000 days, 0.2584 320128 - condition - Essential hypertension - 730 days, 0.2419 4167493 - condition - Pregnancy-induced hypertension - 10000 days, 0.1925 312648 - condition - Benign essential hypertension - 10000 days, 0.1116 433536 - condition - Severe pre-eclampsia - 10000 days, 0.0950 137940 - condition - Transient hypertension of pregnancy - delivered - 10000 days, 0.0929 439393 - condition - Pre-eclampsia - 10000 days, 0.0899 434005 - condition - Morbid obesity - 10000 days, 0.0820 320128 - condition - Essential hypertension - 180 days, 0.0707 42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum - 180 days, 0.0694 4170137 - condition - Non-suppurative otitis media - 10000 days, 0.0650	4014295 - condition - Single live birth - 10000 days, -0.2009
Procedure	27,635		
Labs	26,815	2212545 - labs - Protein, total, except by refractometry; urine - 10000 days, 0.1662 3001582 - labs - Protein/Creatinine [Mass Ratio] in Urine - 10000 days, 0.0836 2213046 - labs - Tissue examination by KOH slide of samples from skin, hair, or nails for fungi or ectoparasite ova or mites (eg, scabies) - 365 days, 0.0630	2212991 - labs - Culture, presumptive, pathogenic organisms, screening only - 10000 days, -0.1593 2212333 - labs - Ferritin - 10000 days, -0.0786
Drugs	19,550		
Specialty	515		

Table A.7: Top 25 features of LASSO pregnancy risk factor model, g^*, ranked by no complication weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12	Age at end date, 0.1546 8527 - race - White, 0.0465	
Conditions	37,795	4047564 - condition - Routine antenatal care - 730 days, 0.1128 4047564 - condition - Routine antenatal care - 365 days, 0.1008 4014295 - condition - Single live birth - 10000 days, 0.0543	320128 - condition - Essential hypertension - 730 days, -0.1583 320128 - condition - Essential hypertension - 10000 days, -0.1489 433736 - condition - Obesity - 10000 days, -0.1366 438480 - condition - Abnormal glucose tolerance in mother complicating pregnancy, childbirth AND/OR puerperium - 10000 days, -0.1235 4024659 - condition - Gestational diabetes mellitus - 10000 days, -0.1062 312648 - condition - Benign essential hypertension - 10000 days, -0.0970 320128 - condition - Essential hypertension - 180 days, -0.0958 4167493 - condition - Pregnancy-induced hypertension - 10000 days, -0.0939 42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum - 10000 days, -0.0674 439393 - condition - Pre-eclampsia - 10000 days, -0.0570 434005 - condition - Morbid obesity - 10000 days, -0.0498
Procedure	27,635	2101813 - procedure - Neuraxial labor analgesia/anesthesia for planned vaginal delivery (this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor) - 10000 days, 0.0537	
Labs	26,815	2212991 - labs - Culture, presumptive, pathogenic organisms, screening only - 30 days, 0.0588 2212991 - labs - Culture, presumptive, pathogenic organisms, screening only - 10000 days, 0.0428	2212545 - labs - Protein, total, except by refractometry; urine - 10000 days, -0.1220 3017250 - labs - Creatinine [Mass/volume] in Urine - 730 days, -0.0450
Drugs	19,550		
Specialty	515		

Table A.8: Top 25 features of pregnancy risk factor model conditioned on history of diabetes, $g^*_{\mathcal{H}=DB}$, ranked by gestational diabetes weight.

Feature Type	# of Features	Top Features (+)	Top Features (-)
Demographics	12		0 - race - No matching concept, -0.1127
Conditions	37,795	4042728 - condition - Blood glucose abnormal, 0.3706 4024659 - condition - Gestational diabetes mellitus, 0.1588 194696 - condition - Dysmenorrhea, 0.1382 443871 - condition - Gestation period, 38 weeks, 0.1237 4151985 - condition - Lower back injury, 0.0858 27674 - condition - Nausea and vomiting, 0.0846 434480 - condition - Syndrome of infant of diabetic mother, 0.0746 4016042 - condition - Diabetic on oral treatment, 0.0642	443412 - condition - Type 1 diabetes mellitus without complication, -0.2545 26378 - condition - Hyperpituitarism, -0.1436 195867 - condition - Noninflammatory disorder of the vagina, -0.1013 81902 - condition - Urinary tract infectious disease, -0.0777 443732 - condition - Disorder due to type 2 diabetes mellitus, -0.0691
Procedure	27,635	2314318 - procedure - Medical nutrition therapy; initial assessment and intervention, individual, face-to-face with the patient, each 15 minutes, 0.0849	2313814 - procedure - Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report, -0.1138 2514530 - procedure - Preventive medicine counseling and/or risk factor reduction intervention(s) provided to an individual (separate procedure); approximately 15 minutes, -0.0713
Labs	26,815	2212611 - labs - Urea nitrogen; quantitative, 0.1217	2212169 - labs - Urinalysis; qualitative or semiquantitative, except immunoassays, -0.0890 2212648 - labs - Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count, -0.0864
Drugs	19,550		
Specialty	515		

Table A.9: Top 25 features of pregnancy risk factor model conditioned on history of diabetes, $g^*_{\mathcal{H}=DB}$, ranked by gestational hypertension weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12		
Conditions	37,795	4170137 - condition - Non-suppurative otitis media, 0.2172 4060157 - condition - Umbilical cord tight around neck - delivered, 0.1502 318443 - condition - Arteriosclerotic vascular disease, 0.1270 4066371 - condition - Cellulitis and abscess of lower leg, 0.1217 376125 - condition - Disorder of eyelid, 0.1153	
Procedure	27,635		2213418 - procedure - Immunization administration (includes percutaneous, intradermal, subcutaneous, or intramuscular injections); 1 vaccine (single or combination vaccine/toxoid), -0.1740 2213244 - procedure - Cytopathology, cervical or vaginal (any reporting system), collected in preservative fluid, automated thin layer preparation; with screening by automated system and manual rescreening or review, under physician supervision, -0.1070
Labs	26,815	2212169 - labs - Urinalysis; qualitative or semiquantitative, except immunoassays, 0.2788 2213094 - labs - Infectious agent antigen detection by immunoassay technique, (eg, enzyme immunoassay [EIA], enzyme-linked immunosorbent assay [ELISA], immunochemiluminometric assay [IMCA]) qualitative or semiquantitative, multiple-step method; Influenza, A or B, each, 0.1902 40763395 - labs - Crystals.amorphous [Presence] in Urine sediment by Light microscopy, 0.1100 2212171 - labs - Urinalysis; microscopic only, 0.0911	2212991 - labs - Culture, presumptive, pathogenic organisms, screening only, -0.1047 3013429 - labs - Basophils [#/volume] in Blood by Automated count, -0.0851
Drugs	19,550		
Specialty	515		

Table A.10: Top 25 features of pregnancy risk factor model conditioned on history of diabetes, $g^*_{\mathcal{H}=DB}$, ranked by no complication weight.

Feature Type	# of Features	Top Features (+)	Top Features (-)
Demographics	12	0 - race - No matching concept, 0.0868	
Conditions	37,795	259848 - condition - Chronic rhinitis, 0.1215 141095 - condition - Acne, 0.1131 437246 - condition - Vitamin B deficiency, 0.1111 138387 - condition - Thyrotoxicosis, 0.1077 78473 - condition - Solitary cyst of breast, 0.1054 257007 - condition - Allergic rhinitis, 0.0945 133727 - condition - Thyrotoxicosis without goiter or other cause, 0.0916	4042728 - condition - Blood glucose abnormal, -0.2158 40443308 - condition - Polycystic ovary syndrome, -0.1746 43531007 - condition - Pre-existing diabetes mellitus in pregnancy, -0.1679 440922 - condition - Diabetic on insulin, -0.1154 27674 - condition - Nausea and vomiting, -0.0967 45757124 - condition - Gestational diabetes mellitus in childbirth, -0.0838
Procedure	27,635	2314285 - procedure - Therapeutic procedure, 1 or more areas, each 15 minutes; neuromuscular reeducation of movement, balance, coordination, kinesthetic sense, posture, and/or proprioception for sitting and/or standing activities, 0.1221 2514530 - procedure - Preventive medicine counseling and/or risk factor reduction intervention(s) provided to an individual (separate procedure); approximately 15 minutes, 0.0999 2211751 - procedure - Ultrasound, pregnant uterus, real time with image documentation, fetal and maternal evaluation plus detailed fetal anatomic examination, transabdominal approach; single or first gestation, 0.0847	2514434 - procedure - Emergency department visit for the evaluation and management of a patient, which requires these 3 key components: An expanded problem focused history; An expanded problem focused examination; and Medical decision making of low complexity. Counseling and/o, -0.1106 2314318 - procedure - Medical nutrition therapy; initial assessment and intervention, individual, face-to-face with the patient, each 15 minutes, -0.1014
Labs	26,815	3031119 - labs - Herpes simplex virus 1+2 IgM Ab [Units/volume] in Serum by Immunoassay, 0.0975 2212168 - labs - Urinalysis, by dip stick or tablet reagent for bilirubin, glucose, hemoglobin, ketones, leukocytes, nitrite, pH, protein, specific gravity, urobilinogen, any number of these constituents; automated, without microscopy, 0.0845	
Drugs	19,550		
Specialty	515		38003845 - specialty - Emergency Medicine, -0.0980

Table A.11: Top 25 features of pregnancy risk factor model conditioned on history of hypertension, $g^*_{\mathcal{H}=HT}$, ranked by gestational diabetes weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12		0 - race - No matching concept, -0.1911 8527 - race - White, -0.1073
Conditions	37,795	4146775 - condition - Incomplete inevitable miscarriage without complication, 0.1363 200843 - condition - Finding of frequency of urination, 0.1101 72711 - condition - Shoulder stiff, 0.1040 40481872 - condition - Multigravida of advanced maternal age, 0.1001 197031 - condition - Intrauterine pregnancy, 0.0963 4302739 - condition - Thigh pain, 0.0911 4195780 - condition - Traumatic dislocation of joint of cervical vertebra, 0.0909	320128 - condition - Essential hypertension, -0.1841 312648 - condition - Benign essential hypertension, -0.1066 4047564 - condition - Routine antenatal care, -0.1012
Procedure	27,635	2753383 - procedure - Resection of Appendix, Percutaneous Endoscopic Approach, 0.1715 2101952 - procedure - Shaving of epidermal or dermal lesion, single lesion, scalp, neck, hands, feet, genitalia; lesion diameter 0.6 to 1.0 cm, 0.1284 2110202 - procedure - Chromotubation of oviduct, including materials, 0.0964	2514527 - procedure - Periodic comprehensive preventive medicine reevaluation and management of an individual including an age and gender appropriate history, examination, counseling/anticipatory guidance/risk factor reduction interventions, and the ordering of laboratory/diag, -0.2761 2313814 - procedure - Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report, -0.1666 2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low, -0.1209
Labs	26,815	2212362 - labs - Glucose; tolerance test (GTT), 3 specimens (includes glucose), 0.1143 3026300 - labs - Glucose [Mass/volume] in Serum or Plasma –2 hours post dose glucose, 0.0997	3026008 - labs - Bacteria identified in Urine by Culture, -0.1025
Drugs	19,550		
Specialty	515		

Table A.12: Top 25 features of pregnancy risk factor model conditioned on history of hypertension, $g^*_{\mathcal{H}=HT}$, ranked by gestational hypertension weight.

Feature Type	# of Features	Top Features (+)	Top Features (-)
Demographics	12		
Conditions	37,795	320128 - condition - Essential hypertension, 0.3342 312648 - condition - Benign essential hypertension, 0.1833 438490 - condition - Severe pre-eclampsia - delivered, 0.1093	381290 - condition - Ocular hypertension, -0.2569 193525 - condition - Abdominal pregnancy, -0.1265 432695 - condition - Post-term pregnancy, -0.1137 4101918 - condition - Noninflammatory disorder of the female genital organs, -0.0974
Procedure	27,635	2212946 - procedure - Blood typing, serologic; Rh (D), 0.0922	2101814 - procedure - Anesthesia for cesarean delivery following neuraxial labor analgesia/anesthesia (List separately in addition to code for primary procedure performed), -0.1065 2110059 - procedure - Destruction of lesion(s), vulva; simple (eg, laser surgery, electrosurgery, cryosurgery, chemosurgery), -0.1011
Labs	26,815	3008598 - labs - Thyroxine (T4) free [Mass/volume] in Serum or Plasma, 0.3029 3007070 - labs - Cholesterol in HDL [Mass/volume] in Serum or Plasma, 0.1590 2213031 - labs - Susceptibility studies, antimicrobial agent; microdilution or agar dilution (minimum inhibitory concentration [MIC] or breakpoint), each multi-antimicrobial, per plate, 0.1074 2212093 - labs - Comprehensive metabolic panel This panel must include the following: Albumin (82040) Bilirubin, total (82247) Calcium, total (82310) Carbon dioxide (bicarbonate) (82374) Chloride (82435) Creatinine (82565) Glucose (82947) Phosphatase, alkaline (84075) Pot, 0.0987	2212345 - labs - Gammaglobulin (immunoglobulin); IgE, -0.1225 3048882 - labs - Streptococcus agalactiae DNA [Presence] in Unspecified specimen by NAA with probe detection, -0.1193 3019897 - labs - Erythrocyte distribution width [Ratio] by Automated count, -0.1147
Drugs	19,550		
Specialty	515		

Table A.13: Top 25 features of pregnancy risk factor model conditioned on history of hypertension, $g^*_{\mathcal{H}=HT}$, ranked by no complication weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12		
Conditions	37,795	381290 - condition - Ocular hypertension, 0.1817 378427 - condition - Tear film insufficiency, 0.1150 4312727 - condition - Secondary physiologic amenorrhea, 0.1121	320128 - condition - Essential hypertension, -0.4051 433736 - condition - Obesity, -0.1721 4167493 - condition - Pregnancy-induced hypertension, -0.1384 432441 - condition - Finding of length of gestation, -0.1198 318800 - condition - Gastroesophageal reflux disease, -0.1099 4145335 - condition - Placental infarct, -0.0912
Procedure	27,635	2211749 - procedure - Ultrasound, pregnant uterus, real time with image documentation, fetal and maternal evaluation, after first trimester (> or = 14 weeks 0 days), transabdominal approach; single or first gestation, 0.1231	
Labs	26,815	2212802 - labs - Inhibin A, 0.1499	3037121 - labs - Protein [Mass/volume] in Urine, -0.1621 3048599 - labs - History of Neural tube defect Narrative, -0.1223 3002109 - labs - Cholesterol in LDL/Cholesterol in HDL [Mass Ratio] in Serum or Plasma, -0.0980 2212227 - labs - Bilirubin; direct, -0.0921 2212545 - labs - Protein, total, except by refractometry; urine, -0.0917
Drugs	19,550		
Specialty	515		

Table A.14: Top 25 features of pregnancy risk factor model conditioned on history of diabetes and hypertension, $g^*_{\mathcal{H}=DB+HT}$, ranked by gestational diabetes weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12		0 - race - No matching concept, -0.1232
Conditions	37,795	4307820 - condition - Unplanned pregnancy, 0.2146 43530807 - condition - Allergic disposition, 0.2001 133141 - condition - Tinea pedis, 0.1872 4273250 - condition - Abscess of buttock, 0.1858 433270 - condition - Cord entanglement without compression, 0.1581 4049417 - condition - Vesicular eczema, 0.1428 762297 - condition - Pain in right knee, 0.1373 432441 - condition - Finding of length of gestation, 0.1368 4138760 - condition - Exacerbation of intermittent asthma, 0.1288	193277 - condition - Deliveries by cesarean, -0.1353 141095 - condition - Acne, -0.1315
Procedure	27,635	2211737 - procedure - Ultrasound, soft tissues of head and neck (eg, thyroid, parathyroid, parotid), real time with image documentation, 0.3072 2110316 - procedure - Cesarean delivery only, 0.2608 4120795 - procedure - Surgical removal of impacted tooth, 0.2232 2780477 - procedure - Resection of Prostate, Open Approach, 0.1292	2414398 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o, -0.2021 2100997 - procedure - Anesthesia for intraperitoneal procedures in lower abdomen including laparoscopy; not otherwise specified, -0.1601 2108115 - procedure - Collection of venous blood by venipuncture, -0.1261
Labs	26,815	40762511 - labs - Human papilloma virus 16+18+31+33+35+39+45+51+52+56+58+59+66+68 DNA [Presence] in Cervix by Probe with signal amplification, 0.1348	
Drugs	19,550		
Specialty	515	38004478 - specialty - Geriatric Medicine, 0.2425	

Table A.15: Top 25 features of pregnancy risk factor model conditioned on history of diabetes and hypertension, $g^*_{\mathcal{H}=DB+HT}$, ranked by gestational hypertension weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12		8522 - race - Other Race, -0.2376
Conditions	37,795	437530 - condition - Disorder of lipid metabolism, 0.1235	4307820 - condition - Unplanned pregnancy, -0.2543 432695 - condition - Post-term pregnancy, -0.2456 4062387 - condition - Injury of muscle and tendon at thorax level, -0.1695 43530807 - condition - Allergic disposition, -0.1186 4275423 - condition - Supraventricular tachycardia, -0.1176
Procedure	27,635	2514520 - procedure - Initial comprehensive preventive medicine evaluation and management of an individual including an age and gender appropriate history, examination, counseling/anticipatory guidance/risk factor reduction interventions, and the ordering of laboratory/diagnos, 0.3614	2101931 - procedure - Biopsy of skin, subcutaneous tissue and/or mucous membrane (including simple closure), unless otherwise listed; single lesion, -0.2028 2775777 - procedure - Dilation of Cervix, Via Natural or Artificial Opening, -0.1735 2110314 - procedure - Postpartum care only (separate procedure), -0.1568 2211737 - procedure - Ultrasound, soft tissues of head and neck (eg, thyroid, parathyroid, parotid), real time with image documentation, -0.1314
Labs	26,815	2212300 - labs - Cyanocobalamin (Vitamin B-12), 0.1611 3027144 - labs - Progesterone [Mass/volume] in Serum or Plasma, 0.1283 2212095 - labs - Lipid panel This panel must include the following: Cholesterol, serum, total (82465) Lipoprotein, direct measurement, high density cholesterol (HDL cholesterol) (83718) Triglycerides (84478), 0.1234	3029943 - labs - Horowitz index in Arterial blood, -0.1404 3008598 - labs - Thyroxine (T4) free [Mass/volume] in Serum or Plasma, -0.1347
Drugs	19,550		
Specialty	515		

Table A.16: Top 25 features of pregnancy risk factor model conditioned on history of diabetes and hypertension, $g^*_{\mathcal{H}=DB+HT}$, ranked by no complication weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12	8522 - race - Other Race, 0.1718	
Conditions	37,795	4150062 - condition - Knee pain, 0.2841 441788 - condition - Human papilloma virus infection, 0.1953 40481101 - condition - Erythema of skin, 0.1831 315831 - condition - Chronic pulmonary heart disease, 0.1782 73754 - condition - Restless legs, 0.1363 378135 - condition - Facial nerve disorder, 0.1085	443871 - condition - Gestation period, 38 weeks, -0.3000 436659 - condition - Iron deficiency anemia, -0.1039
Procedure	27,635		2514520 - procedure - Initial comprehensive preventive medicine evaluation and management of an individual including an age and gender appropriate history, examination, counseling/anticipatory guidance/risk factor reduction interventions, and the ordering of laboratory/diagnos, -0.1391 2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low, -0.1278 2514527 - procedure - Periodic comprehensive preventive medicine reevaluation and management of an individual including an age and gender appropriate history, examination, counseling/anticipatory guidance/risk factor reduction interventions, and the ordering of laboratory/diag, -0.1171 2108115 - procedure - Collection of venous blood by venipuncture, -0.1145 2414398 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o, -0.1023
Labs	26,815	3005033 - labs - Fetal Nuchal fold Thickness US, 0.1770 3016502 - labs - Oxygen saturation in Arterial blood, 0.1206	2212188 - labs - Albumin; urine (eg, microalbumin), quantitative, -0.3075 2212093 - labs - Comprehensive metabolic panel This panel must include the following: Albumin (82040) Bilirubin, total (82247) Calcium, total (82310) Carbon dioxide (bicarbonate) (82374) Chloride (82435) Creatinine (82565) Glucose (82947) Phosphatase, alkaline (84075) Pot, -0.1544 2212300 - labs - Cyanocobalamin (Vitamin B-12), -0.1535 2213187 - labs - Infectious agent antigen detection by immunoassay with direct optical observation; Streptococcus, group A, -0.1518 2212603 - labs - Triiodothyronine T3; free, -0.1048
Drugs	19,550		
Specialty	515		

Table A.17: Top 25 features of pregnancy risk factor model conditioned on no prior history of diabetes or hypertension, $g^*_{\mathcal{H}=none}$, ranked by gestational diabetes weight.

Feature Type	# of Features	Top Features (+)	Top Features (-)
Demographics	12	8515 - race - Asian, 0.0809	
Conditions	37,795	433736 - condition - Obesity, 0.1910 438480 - condition - Abnormal glucose tolerance in mother complicating pregnancy, childbirth AND/OR puerperium, 0.1757 4320944 - condition - Cellulitis of toe, 0.0857 138113 - condition - Cyst of thyroid, 0.0827 4132434 - condition - Gestation period, 8 weeks, 0.0779	4047564 - condition - Routine antenatal care, -0.4341 138525 - condition - Pain in limb, -0.1096 36713926 - condition - Somatic dysfunction of thoracic region, -0.0926
Procedure	27,635		2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low, -0.1869 2213418 - procedure - Immunization administration (includes percutaneous, intradermal, subcutaneous, or intramuscular injections); 1 vaccine (single or combination vaccine/toxoid), -0.0984 2414398 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o, -0.0901 2101813 - procedure - Neuraxial labor analgesia/anesthesia for planned vaginal delivery (this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor), -0.0835 2212937 - procedure - Antibody screen, RBC, each serum technique, -0.0745 2110307 - procedure - Routine obstetric care including antepartum care, vaginal delivery (with or without episiotomy, and/or forceps) and postpartum care, -0.0720
Labs	26,815	2212392 - labs - Hemoglobin; glycosylated (A1C), 0.1202 2212363 - labs - Glucose; tolerance test, each additional beyond 3 specimens (List separately in addition to code for primary procedure), 0.0878 3031119 - labs - Herpes simplex virus 1+2 IgM Ab [Units/volume] in Serum by Immunoassay, 0.0773	2212991 - labs - Culture, presumptive, pathogenic organisms, screening only, -0.1671 3016699 - labs - Glucose [Mass/volume] in Serum or Plasma –1 hour post 50 g glucose PO, -0.1444 2212648 - labs - Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count, -0.1102 2212649 - labs - Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count), -0.0879 3022065 - labs - Statement of adequacy [Interpretation] of Cervical or vaginal smear or scraping by Cyto stain, -0.0760
Drugs	19,550		
Specialty	515	38004461 - specialty - Obstetrics/Gynecology, 0.1354	

Table A.18: Top 25 features of pregnancy risk factor model conditioned on no history of diabetes or hypertension, $g^*_{\mathcal{H}=none}$, ranked by gestational hypertension weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12	8516 - race - Black or African American, 0.0764	8522 - race - Other Race, -0.1459 8515 - race - Asian, -0.1411
Conditions	37,795	42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum, 0.0829 441085 - condition - Elderly primigravida, 0.0711	441641 - condition - Delivery normal, -0.1290 4014295 - condition - Single live birth, -0.0903 43530950 - condition - Complication occurring during pregnancy, -0.0769
Procedure	27,635	2110329 - procedure - Induced abortion, by dilation and curettage, 0.0670 2211397 - procedure - Radiologic examination, spine, lumbosacral; 2 or 3 views, 0.0662	2101813 - procedure - Neuraxial labor analgesia/anesthesia for planned vaginal delivery (this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor), -0.1769 2211749 - procedure - Ultrasound, pregnant uterus, real time with image documentation, fetal and maternal evaluation, after first trimester (> or = 14 weeks 0 days), transabdominal approach; single or first gestation, -0.0907 2314331 - procedure - Chiropractic manipulative treatment (CMT); spinal, 3-4 regions, -0.0866 2110315 - procedure - Routine obstetric care including antepartum care, cesarean delivery, and postpartum care, -0.0672
Labs	26,815	2212545 - labs - Protein, total, except by refractometry; urine, 0.1038 3020876 - labs - Protein [Mass/time] in 24 hour Urine, 0.0745 3001008 - labs - Epithelial cells.squamous [#/area] in Urine sediment by Microscopy high power field, 0.0669	2212991 - labs - Culture, presumptive, pathogenic organisms, screening only, -0.1275
Drugs	19,550		
Specialty	515		38004450 - specialty - Anesthesiology, -0.0851

Table A.19: Top 25 features of pregnancy risk factor model conditioned on no history of diabetes or hypertension, $g^*_{\mathcal{H}=none}$, ranked by no complication weight.

Feature Type	**# of Features**	**Top Features (+)**	**Top Features (-)**
Demographics	12	8527 - race - White, 0.1587 0 - race - No matching concept, 0.1332 Age at end date, 0.0672 8522 - race - Other Race, 0.0530	
Conditions	37,795	4047564 - condition - Routine antenatal care, 0.1454	4126571 - condition - Fetal problem, -0.1053 433736 - condition - Obesity, -0.0669 42872398 - condition - Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum, -0.0613 438480 - condition - Abnormal glucose tolerance in mother complicating pregnancy, childbirth AND/OR puerperium, -0.0554 4132434 - condition - Gestation period, 8 weeks, -0.0534 4188598 - condition - High risk pregnancy, -0.0473
Procedure	27,635	2414397 - procedure - Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low, 0.1130 2514436 - procedure - Emergency department visit for the evaluation and management of a patient, which requires these 3 key components: A detailed history; A detailed examination; and Medical decision making of moderate complexity. Counseling and/or coordination of care with o, 0.0509 2101813 - procedure - Neuraxial labor analgesia/anesthesia for planned vaginal delivery (this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor), 0.0478	
Labs	26,815	2212996 - labs - Culture, bacterial; quantitative colony count, urine, 0.3662 3025891 - labs - Pathology report final diagnosis Narrative, 0.1078 3002529 - labs - ABO group [Type] in Blood, 0.0812 2212991 - labs - Culture, presumptive, pathogenic organisms, screening only, 0.0763 3013650 - labs - Neutrophils [#/volume] in Blood by Automated count, 0.0515	3009214 - labs - Lutropin [Units/volume] in Serum or Plasma, -0.0821 2212545 - labs - Protein, total, except by refractometry; urine, -0.0790 3026008 - labs - Bacteria identified in Urine by Culture, -0.0629 3049187 - labs - Glomerular filtration rate/1.73 sq M.predicted among non-blacks [Volume Rate/Area] in Serum, Plasma or Blood by Creatinine-based formula (MDRD), -0.0564 3033543 - labs - Specific gravity of Urine, -0.0521 3051971 - labs - Cytology report of Cervical or vaginal smear or scraping Cyto stain.thin prep, -0.0501
Drugs	19,550		
Specialty	515		

Timeliness of Risk Predictions

We report additional plots in Figure A-2 showing the distribution of the earliest risk predictions for ELASTIC-NET and XGBOOST to show that risk factors are consistently caught early (before gestation) across classification algorithms.

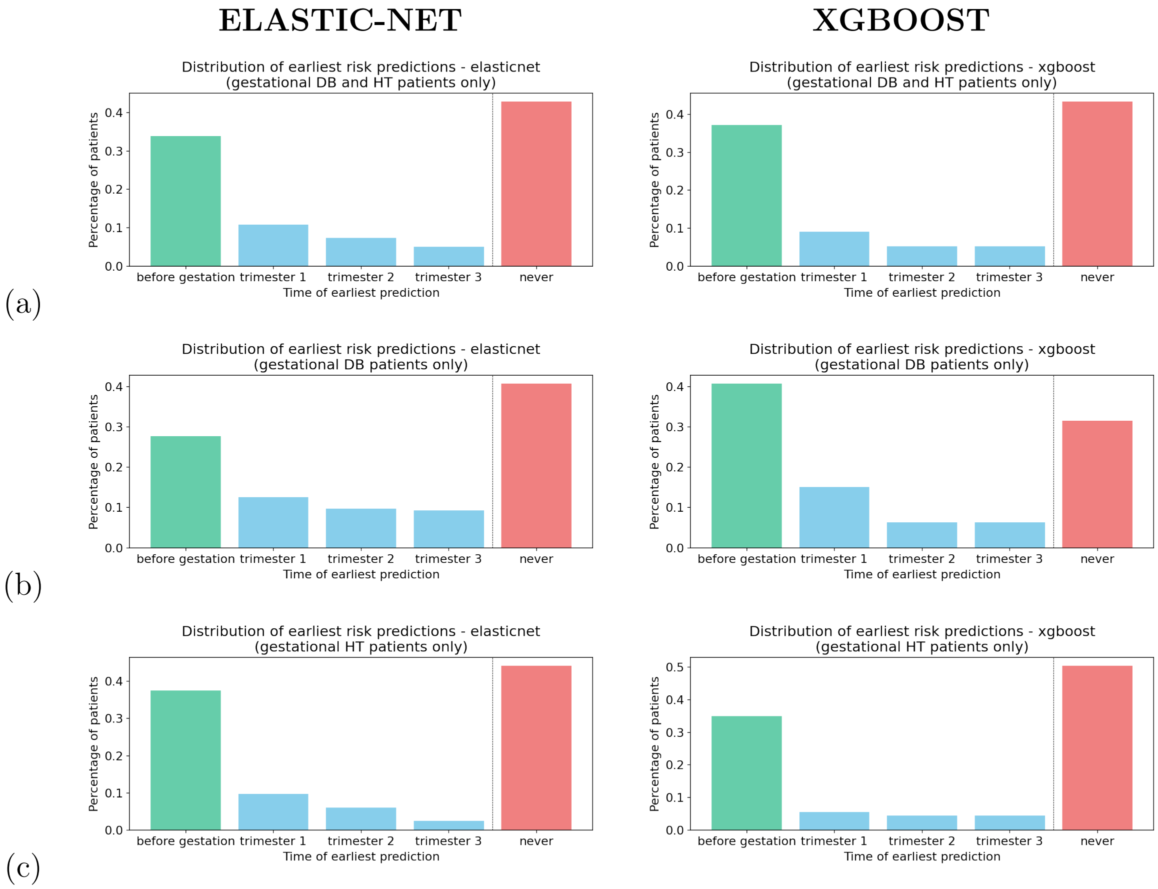


Figure A-2: Distribution of earliest risk predictions for patients at risk of (a) both gestational DB and HT, (b) only gestational DB, and (c) only gestational HT.

A.5 Exemption from Institutional Review Board (IRB) Approval

The user studies for understanding the effect of explanations on prediction of pregnancy and pregnancy risk (Section 6.1 and 6.2) were approved for exemption from Institutional Review Board (IRB) approval by Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects under the following criteria for exemption as defined by Federal regulation 45 CFR 46:

- Exempt Category 3 – Benign Behavioral Intervention, 45 CFR 46.104(d)(3)
- Exempt Category 2 – Educational Testing, Surveys, Interviews or Observation, 45 CFR 46.104(d)(2)

Bibliography

[1] CDC. Unintended Pregnancy. `https://www.cdc.gov/reproductivehealth/contraception/unintendedpregnancy/index.htm`. Accessed: 2022-07-29.

[2] OHDSI. OMOP Common Data Model. `https://ohdsi.github.io/CommonDataModel/index.html`. Accessed: 2022-07-29.

[3] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016.

[4] Trends in Pregnancy and Childbirth Complications in the U.S. `https://www.bcbs.com/the-health-of-america/reports/trends-in-pregnancy-and-childbirth-complications-in-the-us#pre-ex`, june 2020. Accessed:2022-5-10.

[5] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

[6] Elizabeth C. Ailes, Regina M. Simeone, April L. Dawson, Emily E. Petersen, and Suzanne M. Gilboa. Using insurance claims data to identify and estimate critical periods in pregnancy: An application to antidepressants. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 106(11):927–934, November 2016.

[7] Judith W. Alexander and Marlene C. Mackey. Cost Effectiveness of a High-Risk Pregnancy Program. *Care Management Journals*, 1(3):170–174, January 1999.

[8] Marisa L Alunni, Hilary A Roeder, Thomas R Moore, and Gladys A Ramos. First trimester gestational diabetes screening–change in incidence and pharmacotherapy need. *Diabetes research and clinical practice*, 109(1):135–140, 2015.

[9] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai, and the Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, November 2020.

[10] Pierre-Olivier Blotière, Alain Weill, Marie Dalichampt, Cécile Billionnet, Myriam Mezzarobba, Fanny Raguideau, Rosemary Dray-Spira, Mahmoud Zureik, Joël Coste, and François Alla. Development of an algorithm to identify pregnancy episodes and related outcomes in health care claims databases: An application to antiepileptic drug use in 4.9 million pregnant women in France. *Pharmacoepidemiology and Drug Safety*, 27(7):763–770, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.4556.

[11] Thomas Bodenheimer and Rachel Berry-Millett. Care management of patients with complex health care needs. *Research Synthesis Report*, (19):40, December 2009.

[12] Anne F. Carman, Catherine R. Coverston, Rosanne Schwartz, and Myrna L. Warnick. Evaluation of perinatal care management programs: an integrated review. *Care Management Journals: Journal of Case Management ; The Journal of Long Term Home Health Care*, 5(1):19–24, 2004.

[13] Mary C. Carolan-Olah. Educational and intervention programmes for gestational diabetes mellitus (GDM) management: An integrative review. *Collegian*, 23(1):103–114, March 2016.

[14] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N. Balasubramanian. Neural Network Attributions: A Causal Perspective. *arXiv:1902.02302 [cs, stat]*, July 2019. arXiv: 1902.02302.

[15] Marianne Clausel, Thomas Oberlin, and Valérie Perrier. The monogenic synchrosqueezed wavelet transform: a tool for the decomposition/demodulation of AM–FM images. *Applied and Computational Harmonic Analysis*, 39(3):450–486, November 2015.

[16] Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the roc curve. *Advances in neural information processing systems*, 17, 2004.

[17] Radwa El Shawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 275–280, June 2019. ISSN: 2372-9198.

[18] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree, 2017.

[19] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. Technical Report arXiv:1910.06358, arXiv, December 2021. arXiv:1910.06358 [cs, stat] type: article.

[20] Flavio Fuchs. Why Do Black Americans Have Higher Prevalence of Hypertension? | Hypertension.

[21] Yajie Gao, Shuaijun Ren, Huanzhen Zhou, and Rongrong Xuan. Impact of Physical Activity During Pregnancy on Gestational Hypertension. *Physical Activity and Health*, 4(1):32–39, April 2020. Number: 1 Publisher: Ubiquity Press.

[22] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[23] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, 2014.

[24] Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740, 2016.

[25] A. Hamza, D. Herr, E. F. Solomayer, and G. Meyberg-Solomayer. Polyhydramnios: Causes, Diagnosis and Therapy. *Geburtshilfe und Frauenheilkunde*, 73(12):1241–1246, December 2013.

[26] Valerie A. Holmes, Ian S. Young, Christopher C. Patterson, Donald W.M. Pearson, James D. Walker, Michael J.A. Maresh, David R. McCance, and for the Diabetes and Pre-eclampsia Intervention Trial Study Group. Optimal Glycemic Control, Pre-eclampsia, and Gestational Hypertension in Women With Type 1 Diabetes in the Diabetes and Pre-eclampsia Intervention Trial. *Diabetes Care*, 34(8):1683–1688, July 2011.

[27] Clemens S. Hong Hong. Caring for High-Need, High-Cost Patients: What Makes for a Successful Care Management Program? Technical report, Commonwealth Fund, New York, NY United States, August 2014.

[28] Elham Hosseini, Mohsen Janghorbani, and Zahra Shahshahan. Comparison of risk factors and pregnancy outcomes of gestational diabetes mellitus diagnosed during early and late pregnancy. *Midwifery*, 66:64–69, 2018.

[29] M.D. Janice L. Henderson. High-risk pregnancy: What you need to know.

[30] Aria Khademi and Vasant Honavar. A Causal Lens for Peeking into Black Box Predictive Models: Predictive Model Interpretation via Causal Attribution. *arXiv:2008.00357 [cs, stat]*, August 2020. arXiv: 2008.00357.

[31] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684, San Francisco California USA, August 2016. ACM.

[32] Zohra S. Lassi, Tarab Mansoor, Rehana A. Salam, Jai K. Das, and Zulfiqar A. Bhutta. Essential pre-pregnancy and pregnancy interventions for improved maternal, newborn and child health. *Reproductive Health*, 11(1):S2, August 2014.

[33] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H. Tajmir, Claude E. Guerrier, Sarah A. Ebert, Stuart R. Pomerantz, Javier M. Romero, Shahmir Kamalian, Ramon G. Gonzalez, Michael H. Lev, and Synho Do. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 3(3):173–182, March 2019. Number: 3 Publisher: Nature Publishing Group.

[34] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3), Sep 2015.

[35] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *CoRR*, abs/1601.07996, 2016.

[36] Scott Lundberg, Bala Nair, Monica Vavilala, Mayumi Horibe, Michael Eisses, Trevor Adams, David Liston, Daniel Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. 10 2017.

[37] Sarah C. MacDonald, Jacqueline M. Cohen, Alice Panchaud, Thomas F. McElrath, Krista F. Huybrechts, and Sonia Hernández-Díaz. Identifying pregnancies in insurance claims data: Methods and application to retinoid teratogenic surveillance. *Pharmacoepidemiology and Drug Safety*, 28(9):1211–1221, September 2019.

[38] Amy Matcho, Patrick Ryan, Daniel Fife, Dina Gifkins, Chris Knoll, and Andrew Friedman. Inferring pregnancy episodes and outcomes within a network of observational databases. *PLoS ONE*, 13(2):e0192033, February 2018.

[39] Julie A. Meek. Predictive Modeling and Proactive Care Management: Part 1. *Professional Case Management*, 8(4):170–174, August 2003.

[40] Lisa E Moore. Amount of polyhydramnios attributable to diabetes may be less than previously reported. *World Journal of Diabetes*, 8(1):7–10, January 2017.

[41] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning – problems, methods and evaluation, 2020.

[42] Shane T Mueller, Elizabeth S Veinott, Robert R Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J Clancey. Principles of Explanation in Human-AI Systems. page 10.

[43] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[44] Emily E. Petersen. Racial/Ethnic Disparities in Pregnancy-Related Deaths — United States, 2007–2016. *MMWR. Morbidity and Mortality Weekly Report*, 68, 2019.

[45] Jennifer Podulka, M P Aff, Elizabeth Stranges, and Claudia Steiner. Hospitalizations Related to Childbirth, 2008. page 10.

[46] Lore Raets, Kaat Beunen, and Katrien Benhalima. Screening for Gestational Diabetes Mellitus in Early Pregnancy: What Is the Evidence? *Journal of Clinical Medicine*, 10(6):1257, January 2021. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[48] J. A. Rowan, A. Budden, V. Ivanova, R. C. Hughes, and L. C. Sadler. Women with an HbA1c of 41–49 mmol/mol (5.9–6.6%): a higher risk subgroup that may benefit from early pregnancy intervention. *Diabetic Medicine*, 33(1):25–31, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/dme.12812.

[49] Tania Schink, Nadine Wentzell, Katarina Dathe, Marlies Onken, and Ulrike Haug. Estimating the Beginning of Pregnancy in German Claims Data: Development of an Algorithm With a Focus on the Expected Delivery Date. *Frontiers in Public Health*, 8, 2020.

[50] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.

[51] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[52] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5), September 2020. arXiv: 2002.08596.

[53] Arianne N Sweeting, Glynis P Ross, Jon Hyett, Lynda Molyneaux, Maria Constantino, Anna Jane Harding, and Jencia Wong. Gestational diabetes mellitus in early pregnancy: evidence for poor pregnancy outcomes despite treatment. *Diabetes care*, 39(1):75–81, 2016.

[54] Pang-Ning Tan. Introduction to data mining. 2018.

[55] Helena J. Teede, Cheryce L. Harrison, Wan T. Teh, Eldho Paul, and Carolyn A. Allan. Gestational diabetes: Development of an early risk prediction tool to facilitate opportunities for prevention. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 51(6):499–504, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1479-828X.2011.01356.x.

[56] Mitsumasa Umesawa and Gen Kobashi. Epidemiology of hypertensive disorders in pregnancy: prevalence, risk factors, predictors and prognosis. *Hypertension Research*, 40(3):213–220, March 2017.

[57] Josien A Terwisscha Van Scheltinga, Ineke Krabbendam, and Marc EA Spaanderman. Differentiating between gestational and chronic hypertension; an explorative study. *Acta Obstetricia et Gynecologica Scandinavica*, 3(92):312–317, 2013.

[58] B. Vasapollo, G. P. Novelli, G. Gagliardi, G. M. Tiralongo, I. Pisani, D. Manfellotto, L. Giannini, and H. Valensise. Medical treatment of early-onset mild gestational hypertension reduces total peripheral vascular resistance and influences maternal and fetal complications. *Ultrasound in Obstetrics & Gynecology*, 40(3):325–331, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/uog.11103.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[60] Tao Zhou, Shan Du, Dianjianyi Sun, Xiang Li, Yoriko Heianza, Gang Hu, Litao Sun, Xiaofang Pei, Xiaoyun Shang, and Lu Qi. Frontiers | Prevalence and Trends in Gestational Diabetes Mellitus Among Women in the United States, 2006–2017: A Population-Based Study.

[61] Yeyi Zhu and Cuilin Zhang. Prevalence of Gestational Diabetes and Risk of Progression to Type 2 Diabetes: a Global Perspective. *Current diabetes reports*, 16(1):7, January 2016.