Quantifying Preterm Birth Risk and Heterogeneity Using Evolutionary History and Electronic Health Records

By

Abin Abraham

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of
DOCTOR OF PHILOSOPHY
in
Human Genetics
May 31, 2021
Nashville, Tennessee

Approved:
John A. Capra, Ph.D.
Antonis Rokas, Ph.D.
David Aronoff, M.D.
Digna Velez-Edwards, Ph.D.
Lea Davis, Ph.D.
Ge Zhang, M.D. Ph.D.

`

Dedication:

I dedicate this work to my family and my extended family. Mom and Dad, the many sacrifices you made has enabled me to pursue this work. To Jacob, my brother, who had to follow along for the ride and became my trusted confidant.

Acknowledgements

This work is only possible because of the help of many individuals. First, I am grateful for Tony, my thesis advisor. His unwavering mentorship throughout my graduate training nurtured my growth as a scientist. I learned a great deal of technical skills and built my scientific intuition from Tony. In addition to his extraordinary intelligence and creativity that I admire, by the end of my PhD, I was most influenced by his kindness, patience, and integrity.

I owe many thanks to the Capra lab. Mary Lauren, Ling, Laura, Souhrid, David, and Greg, you were the original crew. Most of what I learned my first two years was from listening and asking you all the questions. More importantly, I appreciated your jokes, lab shenanigans, and Tuesday happy hours. Kelia, Sarah, Evonne, and Bian you all sustained the lab spirit as others graduated. I am most grateful for your camaraderie in making lab meetings and doing science what I looked forward to the most during the pandemic of 2020-21.

I was fortunate to work with many collaborators who played an integral role in this work. Dr. Cosmin A. Bejan helped me obtain critical datasets and shared his advice on querying electronic health records. I am grateful for Dr. Abigail L. LaBella, who has been one of my closest collaborators, for her steadfast enthusiasm, energy, and evolutionary insights through the best and worst of times. Dr. Antonis Rokas was not only a key collaborator but also one of my most influential mentors in graduate school. In addition to his scientific brilliance, Dr. Rokas's presence was a steadying influence through many ups and downs of my graduate training. I am deeply grateful for his support, wisdom, and advocacy for me. I also grateful for Dr. Digna Velez-Edwards, Dr. Lea Davis, Dr. Ge Zhang, and Dr. Dave Aronoff for serving as my committee members, collaborators, and mentors throughout my training.

I thank the Vanderbilt MSTP for being my home during my physician scientist training. The MSTP leadership team is top notch; their work behind the scenes creates the infrastructure and nurtures a collaborative culture that enables our success. Finally, I thank my MSTP cohort, classmates, and, especially Matt W, Matt M, Maxwell, Kelsey, Lizzie, and Tory for being my close companions in think crazy ride!

This work was made possible because of funding support from the Burroughs Wellcome Fund Preterm Birth Initiative, the March of Dimes, the National Institutes of Health (T32GM007347), and the American Heart Association Predoctoral Fellowship (20PRE35080073). The Advanced Computing Center for Research and Education at Vanderbilt University enabled all the analyses described in this work. Many analyses also relied on open-source datasets and computational packages that other scientists made freely available. Open source datasets and code is instrumental in accelerating the progress of science.

Table of Contents

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

1 Introduction

1.1 Preterm birth definition and prevalence

Preterm birth is defined as delivery before 37 weeks of gestation and affects over 15 million babies worldwide every year[1–3]. Preterm birth prevalence varies between 5-18% worldwide[4–6]. In the United States, preterm birth affects between 10-15% of all deliveries with an average of 10% that has remained stable between 2009 to 2019[4,6,7]. Using gestational age, preterm birth can be further stratified based on gestational length with shorter duration indicating more severe preterm birth. According to the World Health Organization, extremely preterm birth occurs before 28 weeks, very preterm occurs before 32 weeks, and late preterm occurs after 32 weeks[4]. The majority of preterm births occurs after 34 weeks (>60%), while extremely and very preterm birth have less than 3% prevalence[8]. Notably, the prevalence of preterm births less than 32 weeks has remained stable over 2007 to 2013 in the United States[9].

1.2 Preterm birth mortality and morbidity

In addition to the high prevalence, preterm birth leads to substantial infant morbidity and mortality. Even though the exact definition of extremely preterm birth varies, birth before 28 weeks contributes disproportionally to infant mortality[10] and, more generally, infant mortality is inversely related to gestational age[11]. Preterm births are the leading cause of infant mortality worldwide[6,12]. Despite improvements in medical interventions, infants born premature have high risk for morbidity. Immediately after delivery, premature infants require prolonged hospitalization and are at high-risk for many complications such as respiratory difficulties, gastrointestinal complications, or intraventricular hemorrhage[12]. Some of these health risks can persist for up to five years of age[13]. Risks of immediate maternal complications and decreased long-term wellbeing after preterm birth have also been documented, especially for cesarean-deliveries[14,15], even though they significantly decrease the odds of perinatal mortality[16]. Reflecting the increased medical care arising from immediate and long-term complications, the total economic burden in the United States for preterm birth was estimated to be around $26.2 billion annually based on an Institute of Medicine review[7].

1.3 Disparity across geography, socio-economic factors, and race

The range of preterm birth prevalence is in part due to disparities between geography, socio-economic factors, and race. In the United States, the southeastern states[17] and rural settings compared to cities have higher prevalence of preterm birth[18]. The geographic patterns likely reflect underlying socio-demographic factors. However, disentangling the influence of these socio-demographic factors on preterm birth is challenging. Preterm birth prevalence, across multiple gestational age windows, is consistently higher in Black women compared to White women[19]. Asian and White women have the lowest rates of preterm birth (~5%), while Black women have twice the risk compared to White women[20]. Low socio-economic and educational status has been robustly linked to increased preterm birth risk[4,21,22] in both Black and White women. Similarly, a neighborhood deprivation index integrating income, education, employment, and housing was associated with preterm birth in White and Black cohorts[23]. While socio-economic factors contribute to preterm birth, they might not, by themselves, explain the observed racial disparities[19]. For example, socio-demographic factors were not associated with very preterm births in Black and White women[19,24]. Moreover, racial disparities persisted even when studying women of similar socio--economic status and similar levels of obstetrical care[25]. The substantial differences in preterm birth risk by race, even after controlling for demographic and socio-economic factors, leaves open a role for genetic contributions to birth timing.

1.4 Risk factors for preterm birth

Many risk factors across maternal, fetal, and placental factors have been associated with preterm birth. Key maternal risk factors include race and demographic variables, as discussed above. Maternal nutritional state, as indexed by a low pre-pregnancy BMI is associated with increased rates of spontaneous preterm birth. Meanwhile, obesity can protect from premature delivery[26]. Pregnancy history is an important determinant of preterm birth risk. For example, a family history or a previous preterm birth are the top risk factors for a future preterm birth[6]. The recurrence risk varies between 15 to 50% based on the gestational age and number of previous deliveries[27]. A shorter inter-pregnancy interval less than six months also increases preterm birth risk[28].

Multiple obstetric factors are associated with preterm birth risk. Although multiple gestations (twins, triplets, etc.) account for less than 3% of infants, they result in 15-20% of all preterm births, and more than half of all twins are born preterm[4]. During pregnancy, vaginal bleeding, especially when associated with placental abruption or previa[29], is another risk factor for preterm birth. Other obstetric risk factors include poly- or oligo-hydramnios, which can occur due to a variety of causes, are also associated with preterm birth[4,30,31].

Intrauterine infection and inflammation are common comorbidities that account for 25-40% of preterm births[32]. Earlier preterm births tend to have higher rates of chorioamnionitis compared to longer gestational deliveries[33]. Infection of the amniotic cavity is thought to most commonly occur via bacterial ascension from the vagina or cervix[4,32]. Notably, lower genital tract microorganisms are rarely identified in the amniotic cavity prior to membrane rupture[34]. While specific microbial agents have been associated with preterm birth, the role of genital infections, such as bacterial vaginosis, for preterm birth risk remains unclear[4].

Environmental factors such as exposure to toxins have also been extensively studied. Tobacco use during pregnancy can double the risk for preterm birth likely due to effect on fetal growth[35,36]. Cocaine and heroin use has been implicated in higher preterm birth risk in several studies[37,38]. Other environmental exposures including the effect of physical activity has been studied but their impact on preterm birth remains inconclusive[4].

1.5 Genetic basis of preterm birth

As suggested by differences in prevalence by race, preterm birth risk may have a substantial genetic basis. Twin-based studies have estimated the heritability, the proportion of phenotypic variation in birth timing explained by genetic variation, to be between 30-40%[3,39–41]. However, these estimates do not identify the genes and/or genetic variants underlying variation in preterm birth risk. Linkage analyses identified a small number of genes associated with preterm birth, including *SERPINB2*, *PAI-2*, *AR*, and *IL2RG*[42,43]. The advent of genome-wide association studies (GWAS) over the past ten years has offered a more powerful approach to evaluate variation across the human genome for association with preterm birth. The earliest GWAS were conducted on small cohorts and discovered only a few associations and they failed to replicate[44–46]. A landmark study conducted on over 45,000 European women identified four genomic regions associated with preterm birth that replicated in an independent cohort of >8,000 women[47]. This study mapped genomic associations to regions near the *EBF1*, *EEFSEC, AGTR2, and WNT4* genes. Further functional analyses implicated a variant in *WNT4* that modified estrogen receptor binding[47]. A subsequent study using a smaller cohort but incorporating familial trios identified 72 candidate biomarker genes[48]. Despite the leap forward in identifying genomic regions associated with preterm birth, the heritability explained by these regions remains quite low (~1%).

1.6 Biological mechanisms of preterm birth

While preterm birth has a genetic basis, it should be understood in the context of broader biological mechanisms of birth timing. Preterm birth is considered a syndrome[49] with one of many possible dysfunctional pathways that triggers a common downstream parturition pathway that leads to premature delivery[49]. One precursor pathway includes anatomical abnormalities such as cervical insufficiency; other pathways include inflammation, or maternal stress[6,41,49]. Environmental risk factors, include smoking, alcohol, and low body mass index[50,51], could also act through distinct precursor pathway and induce preterm delivery.

Of the many pathways, only intra-uterine infection has been causally linked to spontaneous preterm birth[52]. More generally, inflammation arising from infections or comorbid disease is studied extensively as a distinct etiology. The collection of chemokines, cytokines, and uterotonic/inflammatory lipids (*e.g.,* prostaglandins) that are activate in inflammation are thought to trigger parturition[49]. Additionally, a potential gene-by-environment interaction between variants in the regulatory regions of the *TNF* gene and bacterial vaginosis has been documented[53].

Another biological mechanism could involve maintenance of the decidua. The decidua results from morphological changes to the endometrium in preparation for pregnancy and is present throughout the pregnancy. Premature decidual senescence has been observed in preterm birth but not in women who deliver term[54]. Another abnormality, preterm pre-labor rupture of membranes, can activate clotting proteins such as thrombin. Thrombin is known to stimulate myometrium contractility and remodel the uterine spiral arteries and vascular under perfusion has been observed in placentas of one third of patients with preterm labor[55]. Similar pathophysiological placental features have been observed in pre-eclampsia, a risk factor for preterm birth. The exact mechanism that leads to preterm birth and why only some women have decidual pathologies remain open questions.

Finally, mechanical factors such as uterine overdistention may lead to preterm birth. For example, it is well known that multiple gestations (e.g., twins) are more likely to deliver preterm. Additionally, polyhydramnios (extra amniotic fluid) can also increase preterm birth risk. In non-human animal models, mechanical stretching of the uterine myometrium can increase inflammatory molecules. Thus, abnormal uterine distention may serve as a precursor to preterm birth[49].

1.7 Classification of preterm birth subtypes

In addition to the diverse biological pathways and numerous risk factors for preterm birth, the syndrome presents with substantial phenotypic heterogeneity. Preterm birth is typically stratified based on gestational age, and the classification is often further dichotomized into spontaneous vs. medically indicated preterm birth. When labor occurs spontaneously before 37 weeks of gestation with intact membranes it is referred to as spontaneous or idiopathic preterm birth[56]. If fetal membranes rupture before 37 weeks but before the onset of delivery, then this is diagnosed as preterm prelabor rupture of membranes and is distinct from spontaneous preterm birth. All other deliveries where the labor is induced due to maternal or fetal health conditions are referred to as medically indicated preterm births[6]. Fewer than half of all preterm births (45%) are spontaneous with the remaining being medically-indicated[57]. Both gestational age-based, and clinical presentation are too imprecise to capture the various comorbidities across maternal, fetal and placental factors[58]. Some risk factors for preterm birth are comorbid maternal disease. For example, cervical anatomic abnormalities can substantially increase preterm birth risk (OR=6.9)[30]. Other systemic comorbidities such as mental health disorders also increase preterm birth risk (OR=1.8)[30]. In a large multi-ethnic cross-sectional study of 5,828 preterm birth, 70% of preterm births had at least one maternal or fetal comorbidity[31]. The heterogenous clinical presentation in addition to the diverse molecular pathways involved in preterm birth suggests that this disorder is better considered a syndrome rather than a disease[49].

The phenotypic heterogeneity of preterm birth limits our ability to discover the underlying biological mechanisms of birth timing and identify interventions. Classification schemes based on maternal, fetal, and placental comorbidities for preterm birth seek to isolate distinct pathways of preterm birth[31,58–61]. For example, clustering individuals on comorbidities has identified an association with the insulin gene, even in a small cohort. Using hierarchical clustering on just 1,028 women identified 120 women with familial risk factors where the insulin gene was significantly associated with spontaneous preterm birth[59].

1.8 EHRs capture dense phenotypes well suited for preterm birth

About two decades ago, health care systems began collecting and linking patient DNA to their de-identified electronic health records (EHRs). Subsequently, the electronic medical records and genomics network (eMERGE) was formed in 2007[62]. EHRs capture detailed phenotypic data generated from clinical care. Numerous studies have demonstrated the utility of this rich phenotyped data in uncovering novel genotype-phenotype associations using approaches like PheWAS[63]. Today, databases linking biospecimens to EHRs have exploded in number. One of the largest repositories is the UK Biobank[64]. At Vanderbilt, BioVU is a similar large-scale de-identified electronic health record database linked to genetic data that has amassed over three million electronic records with 100,000+ individuals genotyped.

EHRs present an opportunity to dissect the phenotypic heterogeneity of preterm birth on an unprecedented scale. Compared to traditional studies that require cohort assembly, large databases of EHRs contains millions of individuals from which we can retrospectively identify a specific cohort of interest. EHRs also accumulate phenotypic data longitudinally for each patient. Genetic data linked to de-identified EHRs extend our ability to study the genetic basis of diseases, especially for those with complex phenotypic presentations.

1.9 Evolutionary context of pregnancy

Integrating evolutionary context with the dense phenotypic snapshot of preterm birth in modern populations from EHRs provides another avenue for further characterization of the complexity and drivers of preterm birth. Pregnancy is a defining characteristic of mammalian species. The central role of pregnancy and parturition to the transmission of genetic information inextricably link these traits to evolution. Many pregnancy-related traits are fast evolving. For example, the placenta is a temporary organ that serves to maintain homeostasis between the mother and the fetus[65,66]. Within mammals, the placenta has evolved independently multiple times and demonstrates remarkable morphological variability[67–69]. For humans, understanding the evolutionary pressures on genomic regions associated with adverse pregnancy outcomes, such as a preterm birth, could illuminate genetically-controlled biological pathways for further research.

1.10 Aims of this thesis

In spite of the diverse maternal, fetal, and placental pathologies that can lead to PRETERM BIRTH, family history and prior occurrences remain the leading clinical risk factors. Furthermore, despite the high heritability of preterm birth and difference in prevalence across individuals of different ancestry, its genetic basis remains poorly understood.

In this dissertation, I leverage a large genetic biobank with linked de-identified EHRs to study the phenotypic and evolutionary heterogeneity of preterm birth. In three parts, this dissertation: 1) refines the definition of preterm birth based on associated comorbidities, 2) prioritizes genetic regions associated with preterm birth using an evolutionary perspective, and 3) incorporates genetic and phenotypic predictors from EHRs to develop a machine learning algorithm to predict preterm birth. By studying the heterogeneity of preterm birth, this work has broad impact from informing future investigations into specific etiologies to developing tools to improve patient care.

Refining the definition of preterm birth leads to precise phenotyping that can isolate specific etiologies. In chapter two, I developed an automated phenotyping approach using an unsupervised method called tensor decomposition to identify several sub-phenotypes of preterm birth. The benefits of this approach are two-fold: 1) precise phenotype definitions with distinct longitudinal and comorbid patterns and 2) the ability to be deployed on large databases of EHRs to quickly increase the sample size for genome-wide association studies. High-throughput and effective ascertainment will advance our understanding of preterm birth biology, reveal comorbid patterns, and lead to targeted therapeutic strategies for different types of preterm birth.

As a complementary approach to uncovering the phenotypic heterogeneity, in chapter three I leveraged an evolutionary perspective to identify different evolutionary forces that have shaped genomic regions associated with preterm birth. Notably, this approach uses existing preterm birth associations and does not require new genome-wide association studies. Preterm birth has significant disparities across race, but limited genomic datasets in non-European populations and some evolutionary forces have been shown to drive population-specific genetic differentiation in a number of different complex traits[70]. Thus, investigating the evolutionary dynamics of preterm birth regions can inform genetically-driven differences in population-specific disease risk. Additionally, building evolutionary priors based on signatures of selection on genomic regions will enable prioritization of genomic candidates for further *in vitro* or *in vivo* validation.

In chapter four, I combined comorbid factors with genetic risk for preterm birth to develop a predictive algorithm. This tool has the potential to support clinical decision making and improve maternal health. Existing tools to predict preterm birth have limited accuracy and are applied only in specific clinical contexts[71–73]. I developed a powerful machine learning algorithm to predict preterm birth by incorporating diverse data types from EHRs. Machine learning algorithms can be deployed without burdensome economic costs and generate risk scores relatively quickly at the point of care. These models use billing codes which are commonly used across different health systems, and I show that their strong performance generalizes across distinct health systems. This predictive algorithm is a first step at demonstrating the potential of predictive models using data derived from EHRs.

In summary, this dissertation advances our knowledge of preterm birth heterogeneity and demonstrates the translational potential of predictive algorithms derived from genetic biobanks linked to EHRs. The refined definition of preterm birth and prioritized candidate genomic regions opens the potential for mechanistic insights into birth timing and predictive algorithms that model disease heterogeneity demonstrates the potential for translational clinical impact.

CHAPTER II

2 Resolving the phenotypic heterogeneity of preterm birth

2.1 Introduction

Preterm birth affects approximately 10% of pregnancies and is the leading cause of infant mortality worldwide[4,6,7,74]. The precise mechanisms of birth timing remain poorly understood[6,49]. Many lines of evidence suggest a strong genetic basis for birth timing. For example, family history and a previous preterm birth are the strongest risk increasing factors for preterm birth[4]. Additionally, twin based studies estimate heritability up to 40% [39–41]. Despite evidence for a strong genetic basis, genome-wide association studies (GWASs) to date have had limited success in identifying associated genomic regions[47,48]. The largest and most robust GWAS to date replicated only four regions associated with preterm birth[47].

For genetic and mechanistic studies, the phenotypic heterogeneity of preterm birth[31,49] can limit the power to detect distinct etiologies[58,61]. Preterm birth is often divided into those with spontaneous onset of labor or medically-indicated deliveries[57,61]. This dichotomy does not incorporate the varied clinical presentation and comorbidities that influence preterm birth risk. A majority of preterm birth cases present with at least one complicating maternal/fetal condition[31]. In addition to obstetric factors, chronic and systemic comorbidities such as diabetes or mental health disorders, also increase the risk for preterm birth[30].

A refined definition of preterm birth phenotype has the potential to identify specific disease mechanisms. For example, the etiology of some preterm birth cases can be narrowed down to cervical remodeling that is necessary for parturition[75]. Even in a small cohort (n~100), hierarchical clustering on obstetric variables identified a subset of preterm birth cases associated with the *INS* (insulin) gene[59]. EHRs are a powerful resource for parsing the phenotypic complexity of preterm birth. EHRs are a cost-effective and efficient for assembling large cohorts with broadly sampled disease traits across multiple timepoints. Indeed, leveraging disease traits from EHRs have identified novel genotype-phenotype associations[63,76] and resolved the heterogeneity across disease domains.

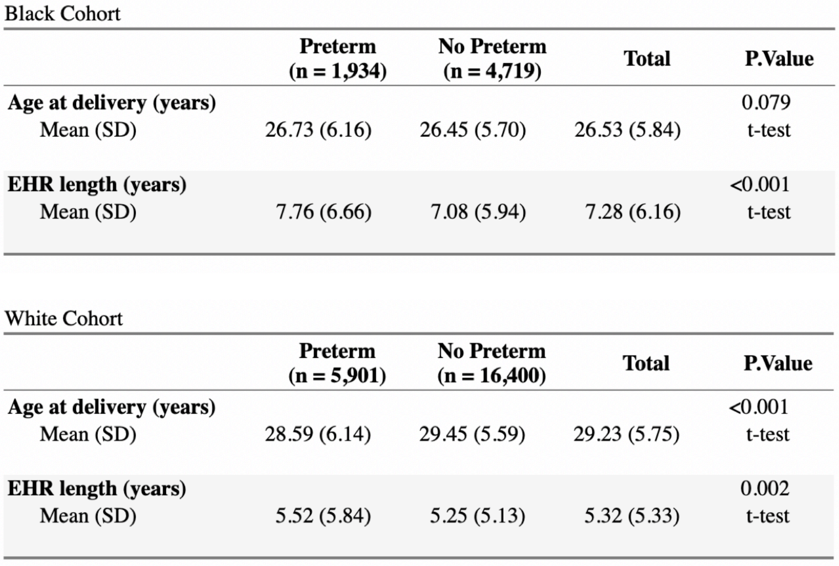
In addition to capturing a broad array of diseases, EHRs record the trajectory of disease over a patient's life. Tensor decomposition is an unsupervised method that can model *ab intio* longitudinal disease trajectories, while simultaneously decomposing a phenotype into latent factors based on their comorbidities. Applying tensor decomposition to EHR data for cardiovascular disease, Zhao, *et al*. demonstrated time-evolving trajectories of distinct subtypes associated with differential long-term survival[77]. EHRs are well suited to investigate phenotypic heterogeneity in pregnancy. During pregnancy there is heightened clinical surveillance, measurable endpoints, and rich documentation of the patient's health. Moreover, a patient health history can affect the risk for adverse pregnancy outcomes and adverse outcomes during pregnancy can have long-term health consequences.

In this study, we investigate the phenotypic heterogeneity of preterm birth in a large cohort of White and Black pregnancies using EHRs linked to a genomic biobank. First, we map the phenotypic heterogeneity by testing for diseases associated with preterm birth across the clinical phenome. We next apply tensor decomposition over nine years of EHR data on a cohort of only preterm births in both white and black women. We identify sub-phenotypes represented by nine and 12 latent factors in the White and Black women, respectively. Each latent factor is described by their associated morbidity (phenotypic signatures) and longitudinal trajectories (temporal signature). We hypothesize that an individual's relative membership in a latent factor will be associated with distinct genetic risk in the preterm birth cohort. To test this hypothesis, we regress polygenic risk scores for comorbid traits with each individual's latent factor weights. The genetic risk score for body mass index (BMI), type 1, and type 2 diabetes were statistically significantly associated with specific latent factors.

2.2 Results

2.2.1 Pregnancy cohort characteristics

We assembled a pregnancy cohort from the Vanderbilt EHR database (>3.2 million records) by identifying women with at least one delivery (n=22,301 White, 6,653 Black, **Error! Reference source not found.**). We ascertained the delivery type (preterm vs. not-preterm) using delivery-specific billing codes and estimated gestational age when available (Methods). We observed a higher proportion of preterm birth in the White (29.0%) and Black (26.4%) cohorts compared to population prevalence; this is likely due to sampling from a tertiary care health system enriched for preterm cases. Age at earliest recorded delivery (EHR-delivery) was similar between preterm and not-preterm birth cases in both White and Black cohorts (**Error! Reference source not found.**). The mean length of EHRs, defined as time between the earliest and most recent billing code, was longer in women with preterm birth compared to the not-preterm women in both White (0.27 years longer) and Black cohorts (0.68 years longer; **Error! Reference source not found.**).



Black Cohort

	Preterm (n = 1,934)	No Preterm (n = 4,719)	Total	P.Value
Age at delivery (years)				0.079
Mean (SD)	26.73 (6.16)	26.45 (5.70)	26.53 (5.84)	t-test
EHR length (years)				<0.001
Mean (SD)	7.76 (6.66)	7.08 (5.94)	7.28 (6.16)	t-test

White Cohort

	Preterm (n = 5,901)	No Preterm (n = 16,400)	Total	P.Value
Age at delivery (years)				<0.001
Mean (SD)	28.59 (6.14)	29.45 (5.59)	29.23 (5.75)	t-test
EHR length (years)				0.002
Mean (SD)	5.52 (5.84)	5.25 (5.13)	5.32 (5.33)	t-test

Table 2.1: Demographic characteristics of pregnancy cohorts. For women with at least one delivery at the Vanderbilt Hospital, we ascertained their earliest delivery as (preterm vs. no preterm) using billing codes and estimated gestational age

2.2.2 Preterm birth is associated with many traits across disease systems

To build a phenome-wide map of traits associated with preterm birth, we regressed delivery type on each phecode, a mapping from the International Statistical Classification of Diseases and Related Health Problems (ICD-9) billing codes, while adjusting for age at first EHR-delivery and length of EHR. Women with at least one preterm birth were cases and all others were controls. We considered phecodes that pass Bonferroni multiple testing correction thresholds ($p_{black} < 5.2e-5$, $p_{white} < 3.8e-5$) as being associated with PRETERM BIRTH.

In the Black cohort, we tested 926 phecodes of which 28 were associated with preterm birth (Figure 2.1A). Six out of 17 phecode disease chapters had at least one association with preterm birth; the pregnancy chapter had the largest number of associations (n=11; Figure 2.1B). Amniotic cavity abnormalities, which include oligo/poly-hydramnios, premature rupture of membranes, infection of amniotic membranes, or spontaneous/artificial rupture of membranes, had the strongest association (p=3.1e-76, OR=5.6). Other preterm birth associated phecodes included hypertensive disorders (preeclampsia p=8.2e-55, OR=5.3; hypertension complicating pregnancy, p=1.9e-26, OR=2.5), cervical incompetence (p=5.2e-26, OR=13.4), and miscarriage (p=2.31, OR=3.5).

In the non-pregnancy disease chapters, the top associations with increased preterm risk included hypertensive traits (essential hypertension, p=2.3e-24, OR=3.4; hypertensive chronic kidney disease, p=3.5e-6, OR=12.5), diabetes (type 2, p=8.9e-18, OR=3.4; type 1, p=6.1e-11, OR=5.9), and morbid obesity (p=1.2e-7, OR=2.0). Many renal traits were also strongly associated with preterm birth (end stage renal disease, p=1.1e-5, OR=11.1; chronic renal failure, p=4.6e-6, OR=12.1; proteinuria, p=6.1e-6, OR=6.4).

In the larger White cohort, we detected more preterm birth associated phecodes (n=55) compared to the Black cohort (Figure 2.1C, D). The top associations were similar across both cohorts. Anemia during pregnancy (p = 9.4e-6, OR=2.9), infection of the genitourinary tract (p=1.3E-5, OR=2.9), and rhesus isoimmunization (p=3.7e-06, OR=2.9) were associated only in the White cohort. In the non-pregnancy chapters, pneumonia (p=8.8, OR=3.8), mental disorders during/after pregnancy (p = 1.6e-7, OR=1.8), and lupus (p= 7.3e-07, OR=3.2) were only associated with preterm birth in the White cohort.

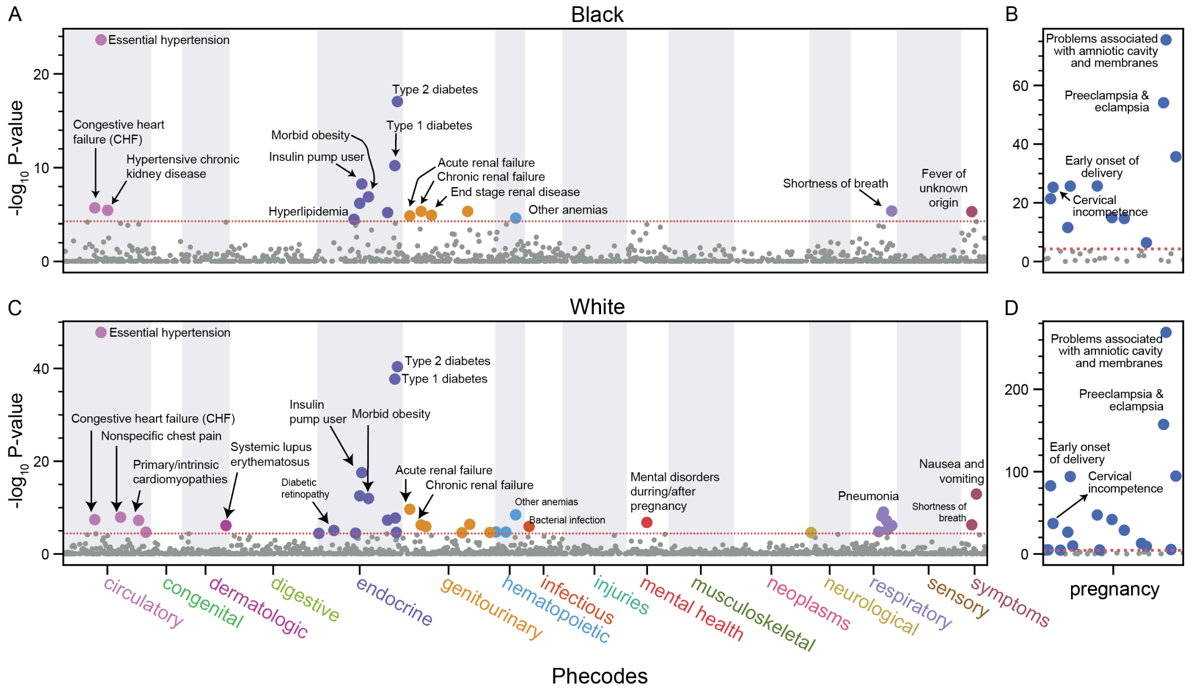


Figure 2.1: Preterm birth is associated with multiple disease phenotypes across the clinical phenome. We tested for an association between preterm birth and disease phenotypes (Phecodes) in a cohort of (A) black (n_cases = 1,934, n_controls = 4,719) and (B) white (n_cases = 5,901, n_controls = 16,400) women while adjusting for age at delivery and the length of EHR. Cases included women with at least one preterm birth while controls included women with term or post-term deliveries. Phenotypes are organized by disease chapter (x-axis) with the pregnancy chapter plotted separately (B, D) and the negative log10 P-value (y-axis). Diseases were considered associated if they passed a Bonferroni correction for number of traits tested (red dotted line) within each cohort (p_black < 5.2e-5, p_white< 3.8e-5)

2.2.3 Tensor decomposition enables discovery of longitudinal sub-phenotypes

The strong associations with preterm birth for distinct traits across multiple systems supports the syndromic nature of preterm birth and suggests that defining preterm birth as a composite of sub-phenotypes based on similar morbidity could improve our understanding of its phenotypic heterogeneity. Thus, we applied tensor decomposition to identify a set of interpretable preterm birth sub-phenotypes with distinct phenotypic and longitudinal signatures (Figure 2.2). In the White and Black preterm birth cohort separately, we constructed a longitudinal disease tensor containing the number of phecodes occurring five years before and after the first delivery for each woman (Figure 2.2A). This tensor has three dimensions: phecodes, time since delivery, and individuals (Figure 2.2B). Next, we factorized the longitudinal disease tensor as an approximate sum of rank-one tensors. Each rank-one tensor represents one latent factor derived from the outer product of three vectors (Figure 2.2E). Each vector for a latent factor quantifies the weight of elements along the phecode, time since delivery, and individual axes. For downstream analyses, we reorganized the weights into matrices by concatenating vectors across latent factors for each of the three tensor axes (Figure 2.2D).

To select the number of latent factors to consider, we defined optimization criteria on the resulting factorizations with the goal of capturing similar morbidity patterns and maximizing interpretability. We performed multiple tensor decompositions with three to 30 latent factors and evaluated the 'between factor uniqueness' (Jaccard Index) and 'within factor coherence (UMass) within the phecode dimension (Methods). After nine and 12 latent factors, the Jaccard index decreased while UMass generally increased for the White and Black cohorts respectively (Figure 2.3). As expected, the sum squared error decreased with more latent factors in both cohorts (Figure 2.3). We selected the number of latent factors to meet the following criteria: minimize sum squared error, maximize factor uniqueness, and maximize within factor identity'. Therefore, we selected nine and twelve latent factors for the White and Black cohorts respectively.

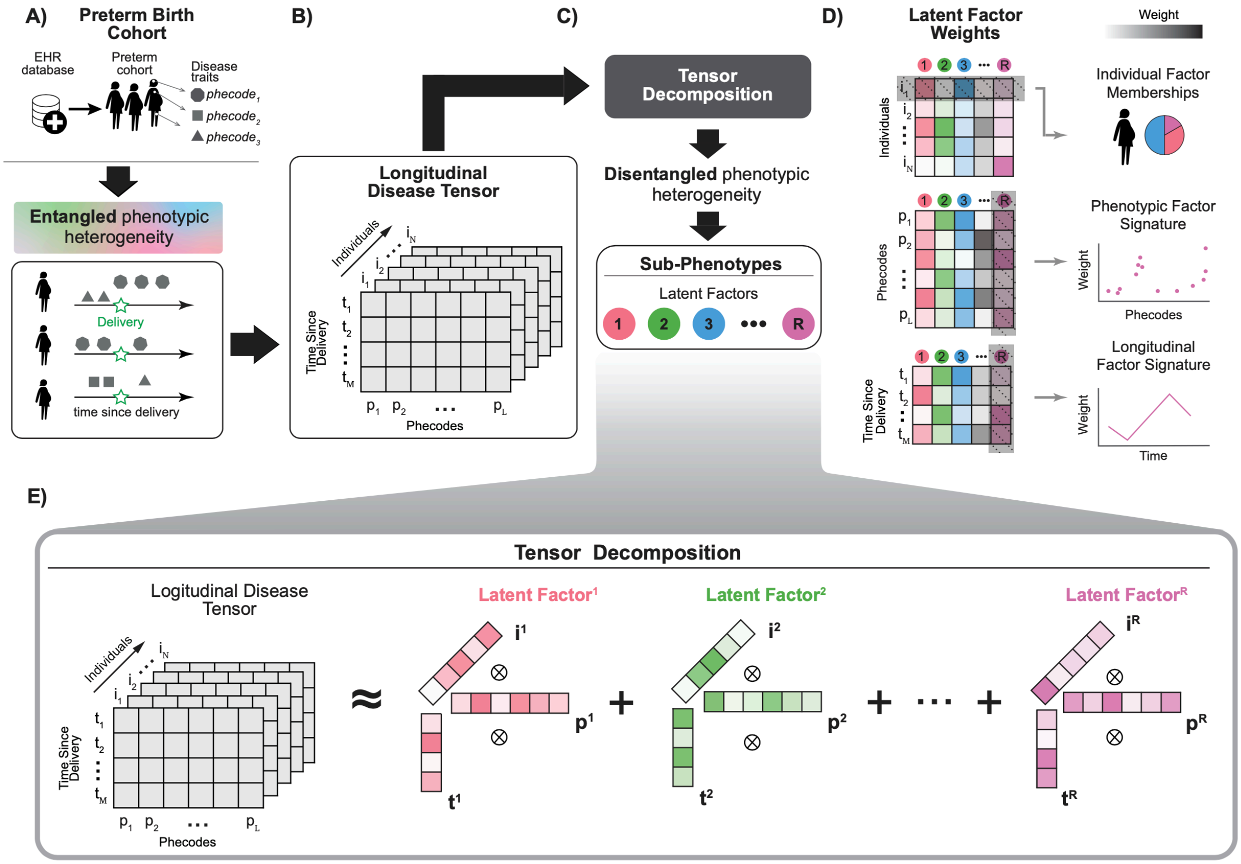


Figure 2.2: Overview of tensor decomposition approach on the preterm birth cohort. A) We considered the cohort of women with at least one preterm birth from our EHR database. B) To capture the phenotypic heterogeneity of preterm birth, we derived phecodes from ICD-9 billing codes occurring up to four years before and after the first recorded delivery. After binning the phecode counts into one-year time intervals, we generated a longitudinal disease tensor with three dimensions: phecodes (x-axis), time since delivery (y-axis), and individuals (z-axis). C) We applied tensor decomposition on the longitudinal disease tensor using parallel factor analysis (PARAFAC) with constraints (Methods). Tensor decomposition was applied separately in the black and white preterm birth cohorts and resulted in a set of latent factors indicative of preterm birth subphenotypes. D) Latent factors consist of weights describing the contribution of elements along the phecode, time since delivery, and individual axes from the original tensor. The weights for an axis across all latent factors are concatenated into weight matrices and analyzed to determine: an individual's relative membership in each latent factor; each latent factor's phenotypic and longitudinal signatures. E) Tensor decomposition using PARFAC approximates the tensor as the outer product of weight vectors along each tensor dimension that are summed across an arbitrary number of R latent factors.

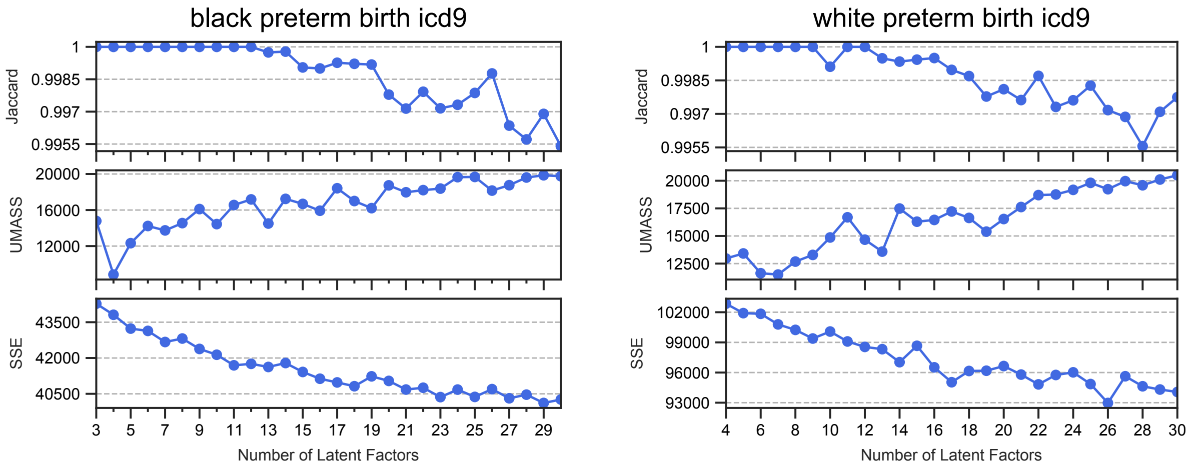


Figure 2.3 Determining the optimal number of latent factors. For the black and white preterm birth cohorts separately, we performed multiple tensor decomposition on the longitudinal disease tensor using three to 30 latent factors. For each decomposition, we measured the Jaccard Index, UMASS (Methods), and sum squared error (SSE).

2.2.4 Latent factors capture distinct phenotypic and longitudinal signatures

To interpret the latent factors after tensor decomposition, we derived a phenotypic signature for each latent factor from the phecode weight matrix for each cohort. The mean weight across each of the phecode chapters revealed distinct signatures for each latent factor in the White cohort (Figure 2.4A). To aid interpretability, we assigned a dominant comorbid axis to the factors based on their highest weights across the phecode chapters (Figure 2.4B). The dominant comorbid axis for factors one to three were in the pregnancy chapter with two and three having smaller weights in the endocrine and circulatory chapters. Factors four and nine had large, localized weights in the mental health and dermatologic chapters respectively. The remaining factors, six and eight, had high weights across multiple chapters reflecting multi-system comorbidity.

We examined latent factors with higher resolution by identifying the top ten phecodes with the highest weights. Latent factor one captures the phecode for preterm birth and other risk factors such as cervical incompetence. Latent factor five has weights distributed between the endocrine and pregnancy chapters and captured different types of diabetes (phecode: 250.1; 250.2, 649.1) and its related comorbidities (diabetic retinopathy, lipid disorders). Similarly, factor four captured related mental health phecodes (anxiety, depression, etc.) and tobacco and substance use disorders.

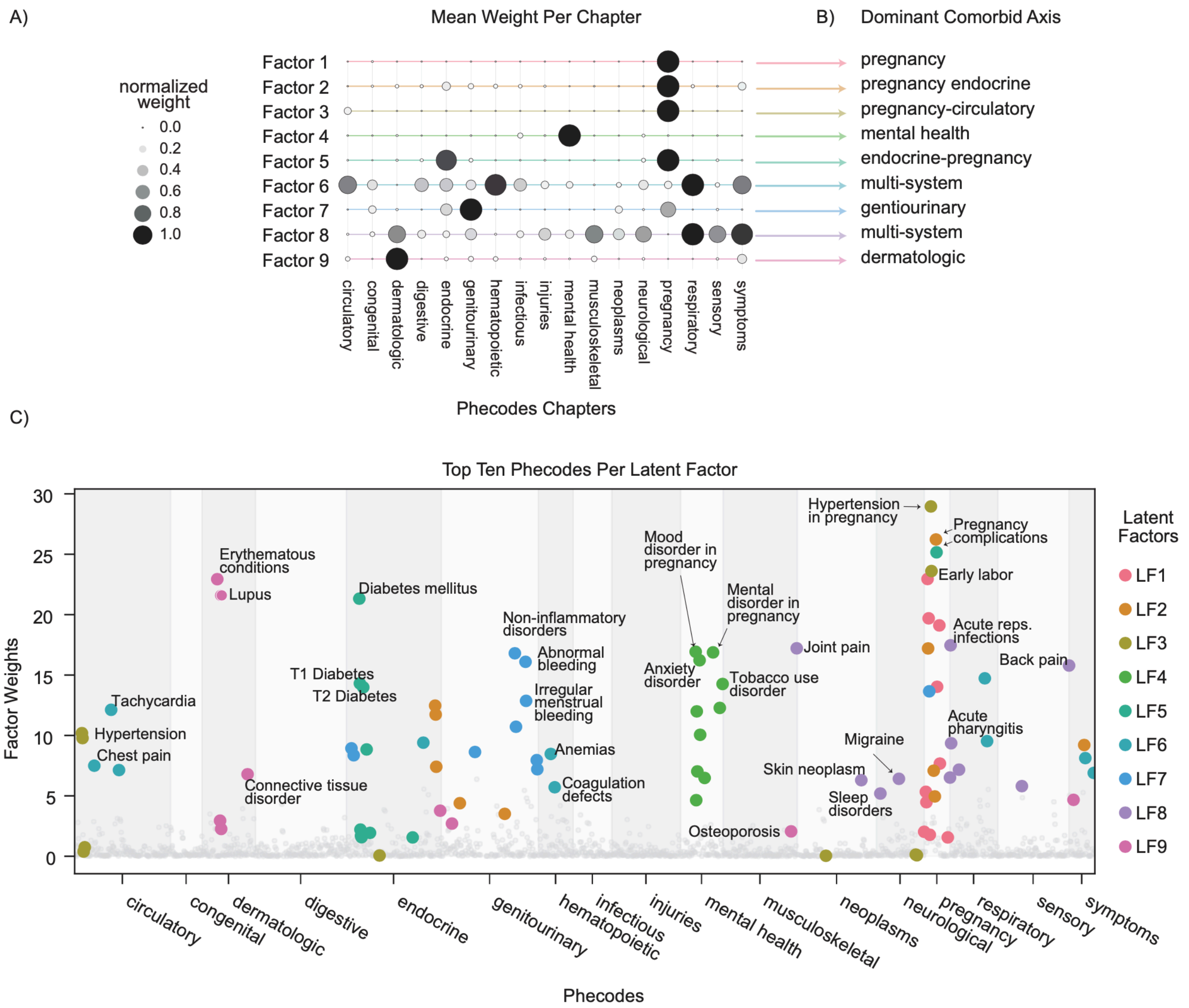


Figure 2.4: Latent factors reveal distinct and interpretable comorbidity signatures. A) After tensor decomposition of the longitudinal disease tensor in the White preterm birth cohort, we analyzed the phecode-by-latent-factor weight matrix. For each factor (rows), we summarized its phenotypic signature based on the mean weight (size) in each phecode chapter (columns). The weights are normalized such that within each factor between zero and one for better interpretability. B) We assigned each latent factor to a dominant comorbid axis based on the largest phenotypic signal(s). C) For each latent factor (colors), we plot the weight (y-axis) for the top ten phecodes with the largest weights. All other phecodes are plotted in gray.

In addition to the phenotypic signatures, we derived the longitudinal trajectories of factors from the time since delivery weight matrix. For each time interval, higher weights indicated a greater morbidity burden. We categorized latent factors into three longitudinal patterns using k-means clustering: peri-pregnancy, chronic-post-pregnancy, and acute pre-pregnancy (Figure 2.5A). The peri-pregnancy group, consisting of factors one, three, four, five, six, and seven, had peak morbidity burden up to one year before delivery. The chronic post-pregnancy group, consisting of factors eight and nine, had sustained morbidity burden up to two years before and four years after delivery. The acute pre-pregnancy group with only factor two had a sharper peak for morbidity burden a year before delivery than the peri-pregnancy group.

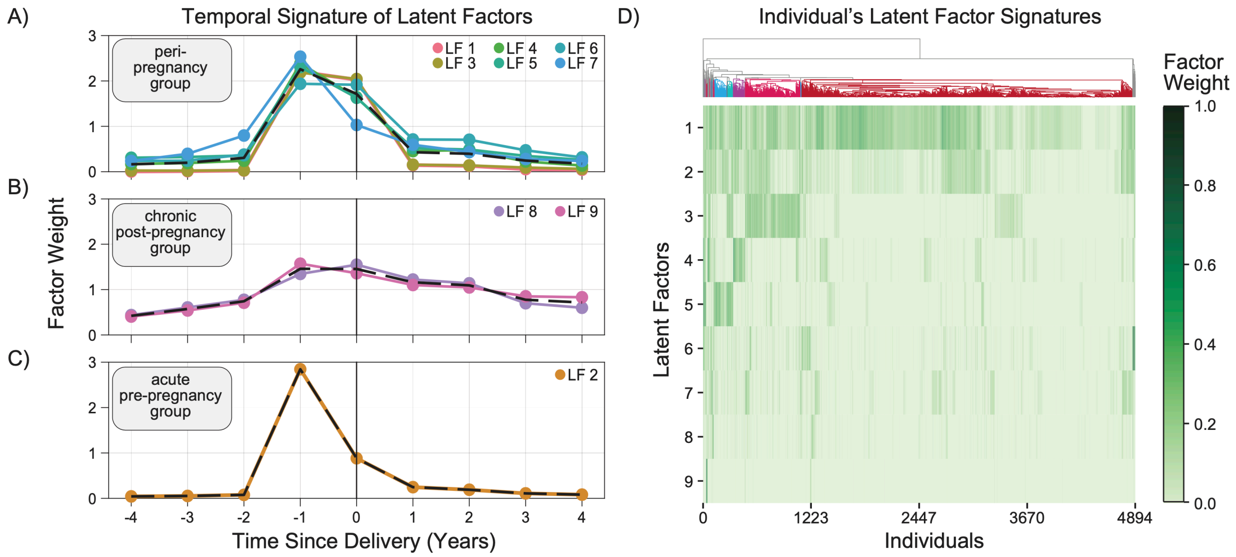


Figure 2.5: Tensor decomposition of white preterm births reveals distinct longitudinal and individual signatures of comorbidity. In the white preterm birth cohort, we used k-means clustering on the nine latent factors based on their weights for the ‘time since delivery’ axis. A, B, C) The temporal signature of each latent factor (colored lines) is captured by its weight (y-axis) across one-year bins before (negative bins) and after delivery (x-axis). All factors have high weights one year before delivery. The centroid of the three groups derived from kmeans clustering is overlaid as the dotted black line. We annotate the three groups based on their temporal pattern: peri-pregnancy, chronic post-pregnancy, and acute pre-pregnancy. D) We also analyzed the individual by latent factor weight matrix to quantify a woman’s (each column) relative membership across factors (rows). Hierarchical clustering revealed subsets of women (dendrogram) with distinct patterns of factor weights. Weight was normalized to a range of one to zero across all factors and individuals.

2.2.5 Individuals predominantly cluster based on latent factor signatures

To explore the role of each factor to different individuals, we examined the ‘individual’ weight matrix across factors in the White preterm birth cohort (Figure 2.5D). An individual’s weight represents their relative membership across the factors. The weights are normalized to the maximum weight in the matrix to aid interpretability. Overall, factors one, two and three, all of which had dominant comorbidities in the pregnancy chapter, had the highest weights across individuals. A subset of individuals had high weights in multiple factors such as factor one and three. The remaining factors had overall lower weights, but distinct and often overlapping weights. Hierarchical clustering of individuals across the latent factors reveals cohort subsets with similar latent factor signatures. After mapping the weights into t-SNE space, individuals separated predominantly based on the latent factor with the highest weight (Figure 2.6A). Some individuals did not cluster based on their dominant latent factor suggesting a mixture of latent factors are associated with their preterm birth phenotype.

Next, we tested whether an individual's weights across latent factors are associated with obstetric factors such as estimated gestational age (EGA) and age at delivery. Latent factors two, three, five, seven and eight had associations ($p<0.05$) with increased EGA by up to four weeks for each unit increase in latent factor weight. Latent factor one ($p<0.05$) associated with decreased EGA by two weeks for each unit increased in latent factor weight. All but latent factor four were associated with age at delivery; only latent factor one had a decrease in age at delivery by 2.5 years for every unit increase in weight; other factors associated with increased age at delivery as high as 12.5 years for each unit increase in weight. (Figure 2.6).

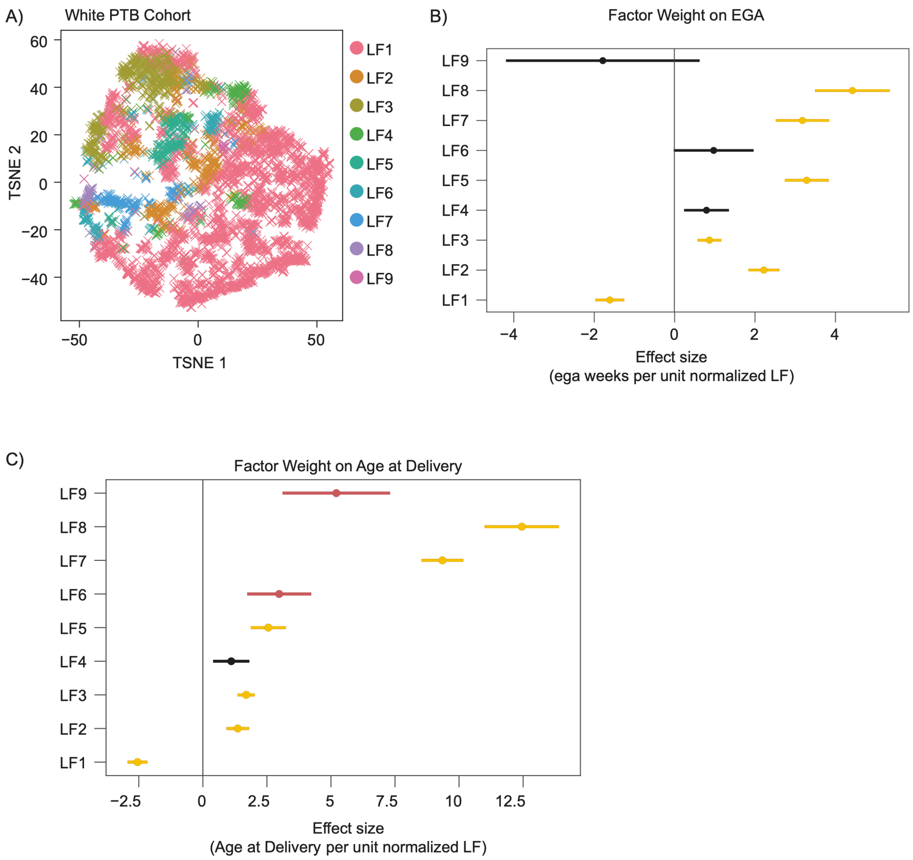


Figure 2.6: Latent factor weights for white women with preterm birth are associated with estimated gestational age and age at delivery. A) After tensor decomposition using nine latent factors, we performed a T-SNE on the weights for each individual across latent factors. Across the first two T-SNE dimensions (x and y axis), each individual (each 'x') is colored based on the latent factor with the highest weight. B) We next regressed estimated gestational age at delivery or C) age at delivery with the normalized weights for each latent factor (y-axis) and plotted the effect size (x-axis). Latent factors that are nominally significant ($p<0.05$) or pass Bonferroni correction for multiple testing across nine latent factors are colored in red and gold respectively.

2.2.6 Black preterm birth cohort exhibits similar latent factor signatures as white preterm cohort

We performed tensor decomposition on Black women with preterm birth and identified twelve latent factors. The phenotype signatures across latent factors had similar patterns to the white preterm birth cohort (Figure 2.7). Several factors had localized weights in specific chapters (pregnancy: factors 1,2; genitourinary: 10,12; dermatologic; 11, mental health: 8). We compared the phenotype weights in the Black cohort to the White cohort and found that most factors had high positive correlations with a small number of factors in the other cohort (Figure 2.7D). This indicated that while the exact factor to disease mapping is arbitrary in each cohort, the underlying morbidity signals are similar to the white preterm birth cohort. We also observed weaker negative correlation in the two cohorts which suggests dissimilarity in phenotype weights between specific pairs of factors (Figure 2.7E). Examining the top phecodes in each latent factor, we observed sickle cell disorder, a prevalent disorder in Black populations, accompanied with anemia-like traits in latent factor nine. Factor 12 also contained disorders highly prevalent in Black women such as polycystic ovaries and menstrual disorders.

After k-means clustering of the longitudinal trajectories, we observed four groups that all exhibited high morbidity weight a year before delivery (Figure 2.7C). The peri/post-pregnancy group, composed of factors five, eight, and ten, had increased morbidity weight after delivery compared to the peri-only pregnancy group (factors one, two, four, and nine). The other two groups exhibited chronic morbidity weights either before and after delivery (‘chronic group’) or only after delivery (chronic post-pregnancy group).

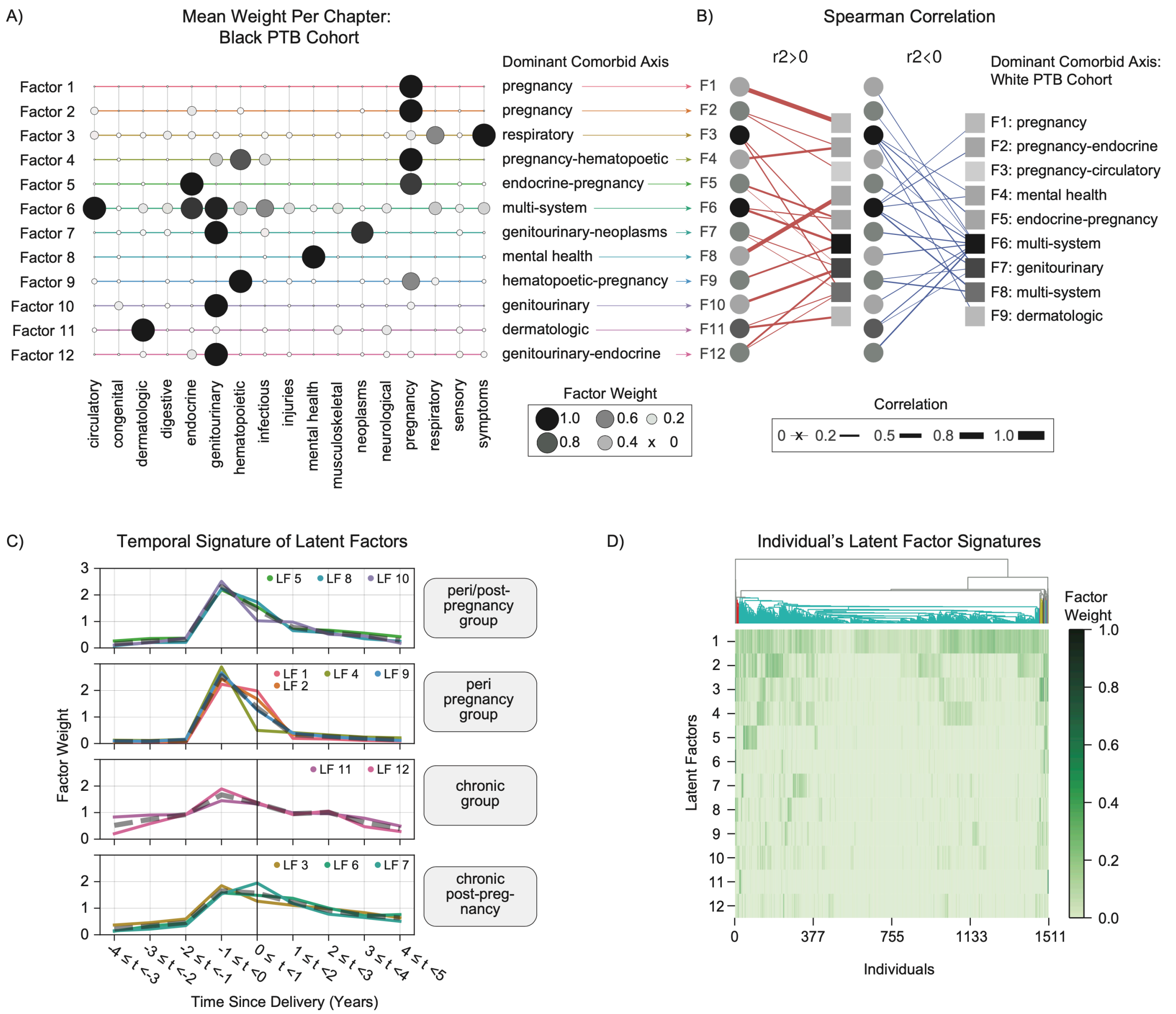


Figure 2.7: Latent factors in a black preterm birth cohort share similar temporal and phenotype profiles with White preterm birth cohort. After tensor decomposition of the longitudinal disease tensor in the Black cohort, we identified 12 latent factors. A) Using the phecode by latent factor weight matrix, we summarized the phenotypic signature of each factor (rows) by the mean weight (size of point) in each phecode chapter (columns). B) The spearman correlation (line thickness) of phecode weights for each latent factor in the white (y-axis) and black (x-axis) preterm birth cohorts. Correlations with p<0.05 are annotated with a star. C) We clustered temporal signature of the latent factors in the black preterm birthcohort as described Figure 4A. Black dashed line represents the centroid of each cluster. D) As described in Figure 4B, the weight of each individual in the Black preterm birth cohort (columns) for each latent factor (rows) demonstrates clustering of women (dendrogram).

Next, we examined the factor membership of individual women. Latent factors one and two, with dominant comorbidities in the pregnancy chapter, had high weights across a large proportion of the cohorts. The remaining factors had distinct subsets of women having higher weights, similar to the White preterm birth cohort. Black women also predominantly clustered according to their factor with the highest weight (Figure 2.8A). Some latent factors in the Black cohort were significantly associated with EGA and age at delivery (Figure 2.8B, C).

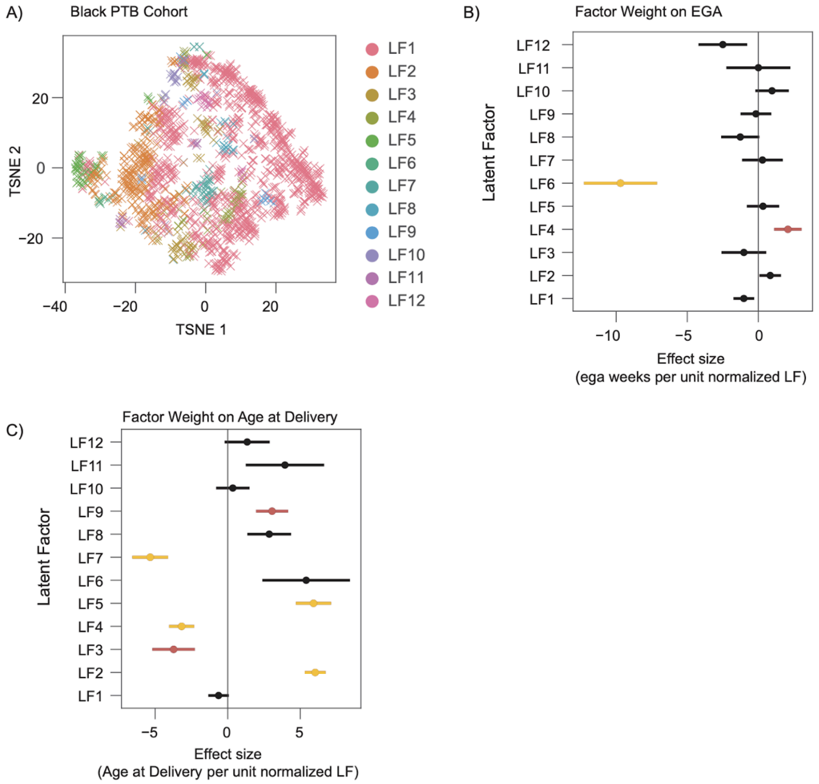


Figure 2.8: Specific latent factors in black women with preterm birth are associated with estimated gestational age and age at delivery. We performed tensor decomposition on a black cohort of women with preterm birth that yielded twelve latent factors. On the weights for each individual across factors, we performed T-SNE and regressed EGA and age at delivery as described in Figure 2.6.

2.2.7 Evaluating latent factors for association with polygenic risk of comorbidities

The latent factors derived from tensor decomposition refined the preterm birth phenotype based on similar morbidity patterns. Therefore, we hypothesized that specific latent factors, driven by their phenotypic signatures, will be associated with the genetic predisposition for preterm birth comorbidities. To capture the genetic predisposition of a trait, we calculated polygenic risk scores using variants weights from previously validated scores from the polygenic risk score catalog. We regressed the PRS on a latent factor weight and adjusted for genetic ancestry. Some traits had multiple PRSs; we corrected for multiple testing (FDR adjusted p value) over all PRS evaluated.

Of all comorbid traits tested, ten had at least one nominally significant association ($p<0.05$) with a latent factor (Figure 2.9). After multiple testing correction, latent factor five that had an endocrine-pregnancy phenotypic signature, associated with type 1 and type 2 diabetes (p.adjusted < 0.05). Similarly, latent factor two with a pregnancy-endocrine phenotypic signature, was associated with type 2 diabetes and BMI (p.adjusted < 0.05).

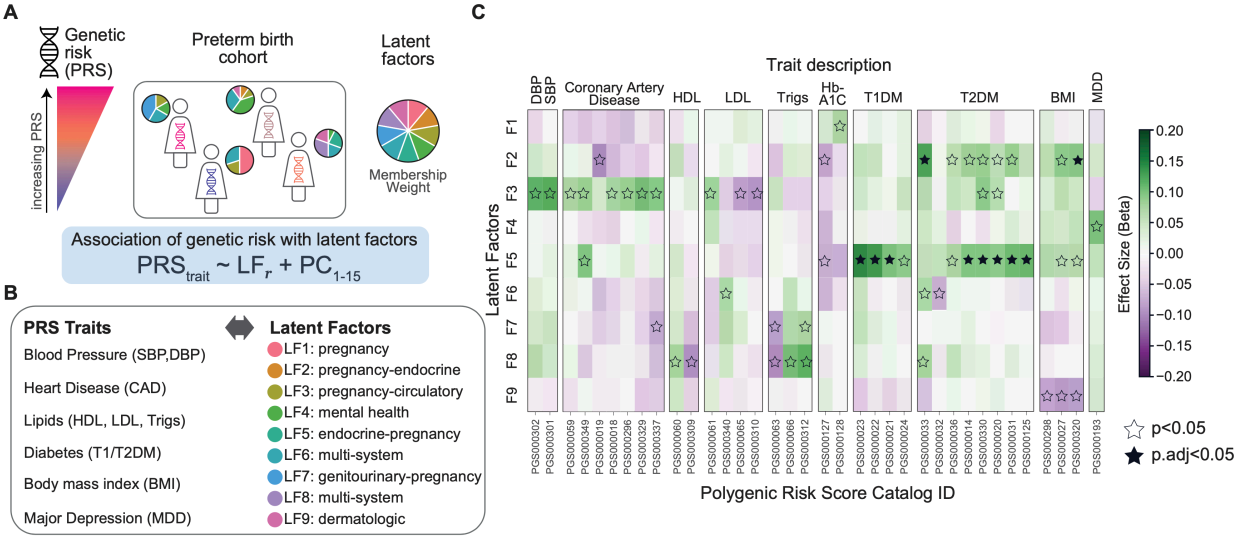


Figure 2.9: Latent factors are associated with polygenic risk score for some preterm birth comorbidities. A) For each woman with preterm birth in the White cohort, we calculated their genetic risk using polygenic risk scores (PRSs) for a variety of comorbidities. We tested for an association between each comorbidity's PRS and the weight for a latent factor and adjusted for 15 genetic ancestry principal components. We repeated this test for latent factors one to nine. B) List of comorbid PRS and their associated latent factor annotated based on its phenotypic signature. C) Heatmap of the effect size (beta coefficient) of the association between each latent factor and each PRS trait. Positive effect size indicates increasing genetic risk with increasing the weight for that latent factor.

2.3 Discussion

Leveraging a large database of de-identified EHRs linked to a genetic biobank, we refined the definition of preterm birth by identifying interpretable sub-phenotypes. We applied tensor decomposition on comorbidities across a patient's EHR, thus incorporating longitudinal trajectories, to yield eight and 12 latent factors in a White and Black cohort of preterm birth deliveries. Each latent factor, representing a sub-phenotype, is distinguished by a temporal signatures five years before and after delivery and the top ten dominant comorbidities. In addition to these interpretable factors, each individual has assigned weights indicating latent factor membership. Specific latent factors were associated with key obstetric variables such as estimated gestational age and age at delivery. To determine if genetic risk for comorbidities is associated with any latent factors, we tested for associations between latent factor weight memberships and the genetic risk of multiple comorbidities. We discover that the genetic risk for BMI and diabetes are significantly associated with specific latent factors in both the White and Black preterm birth cohorts.

Tensor decomposition enables the discovery of latent structure across a large number of commodities while simultaneously capturing longitudinal trajectories. The resulting weights of each latent factor can be mapped directly to specific comorbidities and time intervals. Thus, latent factors are highly interpretable which is a strength of tensor decomposition. Additionally, each individual's factor membership enables association testing with key variables for each latent factor. Tensor decomposition also enables one to apply various constraints that reflect the task at hand. Thus, we applied an orthogonality constraint to each latent factor in the phecode dimension which yielded diverse phenotypic signatures. Although the number of latent factors was different across the Black and White cohorts, some latent factors had very similar phenotypic and temporal signatures.

We tested multiple polygenic risk scores for a given comorbid trait to test for association with latent factors. We obtained polygenic risk scores that were validated in external cohorts from the polygenic risk score catalog. Due to differences in genetic background and differences methods for developing polygenic risk scores, polygenic risk score transferability within population and across populations[78]. Furthermore, assigning a best polygenic risk score is of limited value since demographic and study design factors will influence its performance for out of sample performance. By including multiple polygenic risk scores, we evaluate trends across these variables. Our associations with BMI and diabetes remained robust across multiple polygenic risk scores.

While the use of tensor decomposition is well suited with electronic health record data, there are a few limitations that are important to consider. Electronic records may contain errors in how billing codes are assigned for specific phenotypes. To mitigate this effect, we do not consider extremely rare codes in our cohort. Furthermore, by converting ICD-9 to phecodes, a many to one mapping, an individual phecode will have more support. Another limitation of tensor decomposition is that it is computational expensive. As the input tensor increases, memory requirements may preclude larger cohort from being run. We also observed that while many different constraints could be applied when solving for the tensor decomposition, some of them did not converge on solutions.

2.4 Methods

2.4.1 Ascertaining pregnancy cohort and delivery type from EHRs

From our EHR database (>3.2 million records), we assembled a pregnancy cohort that includes women with at least one delivery at the Vanderbilt University Medical Center (n=35,282). For each woman, we identified if they had multiple pregnancies and ascertained the delivery date and type (preterm vs. not-preterm) using billing codes (ICD-9 or CPT) and estimated gestational age documented in the EHR. First, we group billing codes indicating delivery (delivery-codes) based on their timestamp. We combined delivery-codes into one pregnancy if they occurred within 37 weeks of the most recent delivery-code and repeated until all delivery-codes were assigned to a pregnancy. We grouped EGA values in each woman's EHR into one pregnancy if it was time stamped within the gestational window indicated by the most recent EGA and repeated until all EGA values were assigned to a pregnancy. We determined the delivery type based on the oldest gestational age classification in the pool of delivery-codes and EGA values (i.e. postterm > term > preterm). For each pregnancy, we assign the delivery date as the most recent timestamp from the pool of delivery-codes and EGA values. We have validated this algorithm to identify preterm births (PPV:>90%, Sensitivity: >90%) by chart review as reported previously[79].

2.4.2 Testing disease traits for association with preterm birth

Using self or third-party documented race in the EHR, we filtered the pregnancy cohort into White and Black cohorts. We defined our outcome variable as 'preterm' or 'not-preterm'. Women with at least one preterm birth were 'preterm' (n_black=1,934, n_white=5,901) and all others were 'not-preterm' (n_black=4,719, n_white=16,400). Next, we queried disease traits across the clinical phenome using billing codes. We extracted all ICD-9 billing codes in an individual's EHR and translated them to pheCodes[80]. Phecodes remove unnecessary specificity and collapse ICD-9 codes into clinically relevant groups. Phecodes are further organized into chapters of related diseases. Previous studies have demonstrated that the positive predictive value of a disease increases when multiple counts of that code is required across the EHR[80,81]. Therefore, we binarized phecodes by requiring an individual to have at least four instances of that phecode to be considered positive. We also extracted the age at delivery and the length of an individual's EHR, defined as the time elapsed between the earliest and most recent billing code. Finally, we regressed delivery type on each phecode while adjusting for age at first EHR-delivery and length of EHR. This analysis was performed separately in the black and white cohorts. Only phecodes with at least 100 individuals who were positive were tested for an association with delivery status. To correct for multiple testing, we used a Bonferroni correction over all phecodes tested in each cohort. ICD to phecode translation and association testing was performed using the PheWAS R package[80].

2.4.3 Creating the longitudinal disease tensor from the preterm birth cohort

From the pregnancy cohort, we selected white and black women whose first delivery in the EHR was preterm. For each individual, we binned the number of instances of phecodes into one-year intervals spanning up to four years before and five years after the first EHR-documented preterm delivery. To reduce the size of the tensor, we excluded phecodes occurring in less than 0.5% We assembled a three-dimensional tensor for the white and black cohorts each with the following axes: phecodes, time since delivery, and individuals.

2.4.4 Tensor decomposition using parallel factor analysis with constraints

After generating the longitudinal disease tensor, we performed tensor decomposition using parallel factor analysis using alternating least squares. We required the factorization to be non-negative across all factor dimensions and orthogonal only in the phecode dimension. After fixing the number of latent factors (F), we used alternating least squares for decomposition until model fit converged (R^2≤1e-10). For a given number of latent factors, we repeated the decomposition twenty times with random initializations and retained the model with the highest R^2. To identify the optimum number of latent factors, we performed multiple tensor decompositions with the number of latent factors (F) ranging between three to 30. For each decomposition with F latent factors, we calculated the between factor uniqueness (mean Jaccard Index) and within factor coherence (mean UMass) on the phecode factor matrix.

For each decomposition, we calculated the between phecode-factor independence as the mean Jaccard index of all pairwise factors using weights across the highest fifty phecodes. We also measured the within factor similarity by calculating the UMass metric. UMass[82,83] measures the co-occurrence of similar topics, or phecodes, across a set of EHRs. We calculated the sum of the UMass metric across the top 50 phecodes with the highest weights to obtain a UMass metric per factor. To summarize within factor coherence for a given set of latent factors, we took the mean of the summed UMass metric. To have the Jaccard and mean UMass on the similar scales for comparison, we multiplied UMass metric by negative one and scaled the range between zero and one. Thus, higher scores indicated more average topic coherence.

2.4.5 Factor weights association with estimated gestational age and age at delivery

To test for association between estimated gestational age and age at delivery, we regressed the outcome variable on each latent factor weight. Latent factor weight was normalized within each factor to aid interpretability across factors. A Bonferroni adjusted p-value across all latent factors tested for each outcome variable is used to determine statistical significance.

2.4.6 Calculating and association testing with polygenic risk scores

On the cohort of White women with preterm births, we calculated a polygenic risk score for known comorbidities associated with preterm birth risk. For each of these traits, we downloaded all validated polygenic risk scores (single nucleotide polymorphism, SNP, weights) from the polygenic risk score catalog (PGS)[84]. Many comorbid traits had multiple polygenic risk scores. We calculated the polygenic risk score for an individual as the sum of the number of risk alleles at PRS SNPs weighted by the trait-associated weight downloaded from the PGS. This analysis was performed using the --score function in PLINK v1.90b4s[85]. Pyogenic risk scores are standard normalized to enable interpretation of effect sizes across scores. Multiple testing across latent factors and number of polygenic risk scores was corrected using benjamini-hochberg procedure.

CHAPTER III

3 Evolutionary forces shaping genomic regions associated with preterm birth[1]

3.1 Introduction

Understanding the evolutionary forces that shape variation in genomic regions that contribute to complex traits is a fundamental pursuit in biology. The availability of genome-wide association studies (GWASs) for many different complex human traits[86,87], coupled with advances in measuring evidence for diverse evolutionary forces—including balancing selection[88], positive selection[89], and purifying selection[5] from human population genomic variation data—present the opportunity to comprehensively investigate how evolution has shaped genomic regions associated with complex traits[70,90,91]. However, available approaches for quantifying specific evolutionary signatures are based on diverse inputs and assumptions, and they usually focus on one region at a time. Thus, comprehensively evaluating and comparing the diverse evolutionary forces that may have acted on genomic regions associated with complex traits is challenging. In this study, we develop a framework to test for signatures of diverse evolutionary forces on genomic regions associated with complex genetic traits and illustrate its potential by examining the evolutionary signatures of genomic regions associated with preterm birth , a major disorder of pregnancy. Mammalian pregnancy requires the coordination of multiple maternal and fetal tissues[92,93] and extensive modulation of the maternal immune system so that the genetically distinct fetus is not immunologically rejected[94]. The developmental and immunological complexity of pregnancy, coupled with the extensive morphological diversity of placentas across mammals, suggest that mammalian pregnancy has been shaped by diverse evolutionary forces, including natural selection. In the human lineage, where pregnancy has evolved in concert with unique human adaptations, such as bipedality and enlarged brain size, several evolutionary

[1] *This work has been published in LaBella & Abraham et al.*[262]

hypotheses have been proposed to explain the selective impact of these unique human adaptations on the timing of human birth[95–97]. The extensive interest in the evolution of human pregnancy arises from interest both in understanding the evolution of the human species and also the existence of disorders of pregnancy.

One major disorder of pregnancy is preterm birth, a complex multifactorial syndrome[49] that affects 10% of pregnancies in the United States and more than 15 million pregnancies worldwide each year[2,98]. Preterm birth leads to increased infant mortality rates and significant short- and long-term morbidity[2,4,99]. Risk for preterm birth varies substantially with race, environment, comorbidities, and genetic factors[100]. Preterm birth is broadly classified into iatrogenic preterm birth, when it is associated with medical conditions such as preeclampsia (PE) or intrauterine growth restriction (IUGR), and spontaneous preterm birth (sPTB), which occurs in the absence of preexisting medical conditions or is initiated by preterm premature rupture of membranes[31,57]. The biological pathways contributing to sPTB remain poorly understood[17], but diverse lines of evidence suggest that maternal genetic variation is an important contributor[39,40,101].

The developmental and immunological complexity of human pregnancy and its evolution in concert with unique human adaptations raise the hypothesis that genetic variants associated with birth timing and sPTB have been shaped by diverse evolutionary forces. Consistent with this hypothesis, several immune genes involved in pregnancy have signatures of recent purifying selection[102] while others have signatures of balancing selection. In addition, both birth timing and sPTB risk vary across human populations[103], which suggests that genetic variants associated with these traits may also exhibit population-specific differences. Variants at the progesterone receptor locus associated with sPTB in the East Asian population show evidence of population-specific differentiation driven by positive and balancing selection[90,104]. Since progesterone has been extensively investigated for sPTB prevention[105], these evolutionary insights may have important clinical implications. Although these studies have considerably advanced our understanding of how evolutionary forces have sculpted specific genes involved in human birth timing, the evolutionary forces acting on pregnancy across the human genome have not been systematically evaluated.

The recent availability of sPTB-associated genomic regions from large genome-wide association studies[47] coupled with advances in measuring evidence for diverse evolutionary forces from human population genomic variation data present the opportunity to comprehensively investigate how evolution has shaped sPTB-associated genomic regions. To achieve this, we developed an approach that identifies evolutionary forces that have acted on genomic regions associated with a complex trait and compares them to appropriately matched control regions. Our approach innovates on current methods by evaluating the impact of multiple different evolutionary forces on trait-associated genomic regions while accounting for genomic architecture-based differences in the expected distribution for each of the evolutionary measures. By applying our approach to 215 sPTB-associated genomic regions, we find significant evidence for at least one evolutionary force on 120 regions, and illustrate how this evolutionary information can be integrated into interpretation of functional links to sPTB. Finally, we find enrichment for nearly all of the evolutionary metrics in sPTB-associated regions compared to the genomic background, and for measures of negative selection compared to the matched regions that take into account genomic architecture. These results suggest that a mosaic of evolutionary forces likely influenced human birth timing, and that evolutionary analysis can assist in interpreting the role of specific genomic regions in disease phenotypes.

3.2 Accounting for genomic architecture in evolutionary measures

In this study, we compute diverse evolutionary measures on sPTB-associated genomic regions to infer the action of multiple evolutionary forces (Table 3.1 Evolutionary measures computed on sPTB-associated genomic regions with the corresponding evolutionary signature used to infer the evolutionary force and the associated timescale. GERP: Genomic evolutionary rate profiling. iHS: integrated haplotype score. XP-EHH: cross-population extended haplotype homozygosity (EHH). iES: integrated site-specific EHH. TMRCA: time to most recent common ancestor derived from ARGweaver. Alignment block age was calculated using 100-way multiple sequence alignments to determine the oldest most recent common ancestor for each alignment block.). While various methods to detect signatures of evolutionary forces exist, many of them lack approaches for determining statistically significant observations or rely on the genome-wide background distribution as the null expectation to determine statistical significance (e.g., outlier-based methods)[106,107]. Comparison to the genome-wide background distribution is appropriate in some contexts, but such outlier-based methods do not account for genomic attributes that may influence both the identification of variants of interest and the expected distribution of the evolutionary metrics, leading to false positives. For example, attributes such as minor allele frequency (MAF) and linkage disequilibrium (LD) influence the power to detect both evolutionary signatures[87,108,109] and GWAS associations[86]. Thus, interpretation and comparison of different evolutionary measures is challenging, especially when the regions under study do not reflect the genome-wide background.

Here we develop an approach that derives a matched null distribution accounting for MAF and LD for each evolutionary measure and set of regions (Figure 3.1). We generate 5,000 control region sets, each of which matches the trait-associated regions on these attributes (Methods). Then, to calculate an empirical p-value and z-score for each evolutionary measure and region of interest, we compare the median values of the evolutionary measure for variants in the sPTB-associated genomic region to the same number of variants in the corresponding matched control regions (Figure 3.1A, Methods). This reduces the risk for false positives relative to outlier-based methods and enables the comparison of individual genomic regions across evolutionary measures. In addition to examining selection on individual genomic regions, we can combine these regions into one set and test for the enrichment of evolutionary signatures on all significant sPTB-associated genomic regions. Such enrichment analyses can further increase confidence that statistically significant individual regions are not false positives but rather genuine signatures of evolutionary forces.

In this section, we focus on the evaluation of the significance of evolutionary signatures on individual sPTB-associated regions; in a subsequent section, we extend this approach to evaluate whether the set of sPTB-associated regions as a whole has more evidence for different evolutionary forces compared to background sets.

Measures	**Evolutionary signature**	**Evolutionary force**	**Time scale**
PhyloP	Substitution rate	Positive/negative selection	Across species
PhastCons GERP	Sequence conservation	Negative selection	Across species
LINSIGHT			Across species and human populations
F_{ST}	Population differentiation	Local adaptation	Human populations
iHS XP-EHH iES	Haplotype homozygosity	Positive selection	Human populations
Beta Score	Balanced polymorphisms	Balancing selection	Human populations
Allele Age (TMRCA)	Ancestral recombination graphs/Alignments	Evolutionary origin / Negative selection	Human populations
Alignment block age	Sequence conservation	Evolutionary origin / Negative selection	Across species

Table 3.1 Evolutionary measures computed on sPTB-associated genomic regions with the corresponding evolutionary signature used to infer the evolutionary force and the associated timescale. GERP: Genomic evolutionary rate profiling. iHS: integrated haplotype score. XP-EHH: cross-population extended haplotype homozygosity (EHH). iES: integrated site-specific EHH. TMRCA: time to most recent common ancestor derived from ARGweaver. Alignment block age was calculated using 100-way multiple sequence alignments to determine the oldest most recent common ancestor for each alignment block.

To evaluate the evolutionary forces acting on individual genomic regions associated with sPTB, we identified all variants nominally associated with sPTB ($p<10E-4$) in the largest available GWAS[47] and grouped variants into regions based on high LD ($r^2>0.9$). It is likely that many of these nominally associated variants affect sPTB risk, but did not reach genome-wide significance due to factors limiting the statistical power of the GWAS[47]. Therefore, we assume that many of the variants with sPTB-associations below this nominal threshold contribute to the genetic basis of sPTB. We identified 215 independent sPTB-associated genomic regions, which we refer to by the lead variant (SNP or indel with the lowest p-value in that region).

For each of the 215 sPTB-associated genomic regions, we generated control regions as described above. The match quality per genomic region, defined as the fraction of sPTB variants with a matched variant averaged across all control regions, is ≥ 99.6% for all sPTB-associated genomic regions. The matched null distribution aggregated from the control regions varied substantially between sPTB-associated genomic regions for each evolutionary measure and compared to the unmatched genome-wide background distribution (Figure 3.1b). The sets of sPTB-associated genomic regions that had statistically significant ($p<0.05$) median values for evolutionary measures based on comparison to the unmatched genome-wide distribution were sometimes different that those obtained based on comparison to the matched null distribution. We illustrate this using the F_{ST} between East Asians and Europeans ($F_{ST\text{-}EurEas}$) for four example sPTB-associated regions labeled by the variant with the lowest GWAS p-value. Regions rs4460133 and rs148782293 reached statistical significance for $F_{ST\text{-}EurEas}$ only when compared to genome-wide or matched distribution respectively, but not both (Figure 3.1b, top row). Using either the genome-wide or matched distribution for comparison of $F_{ST\text{-}EurEas}$, sPTB-associated region rs3897712 reached statistical significance while rs4853012 was not statistically significant.

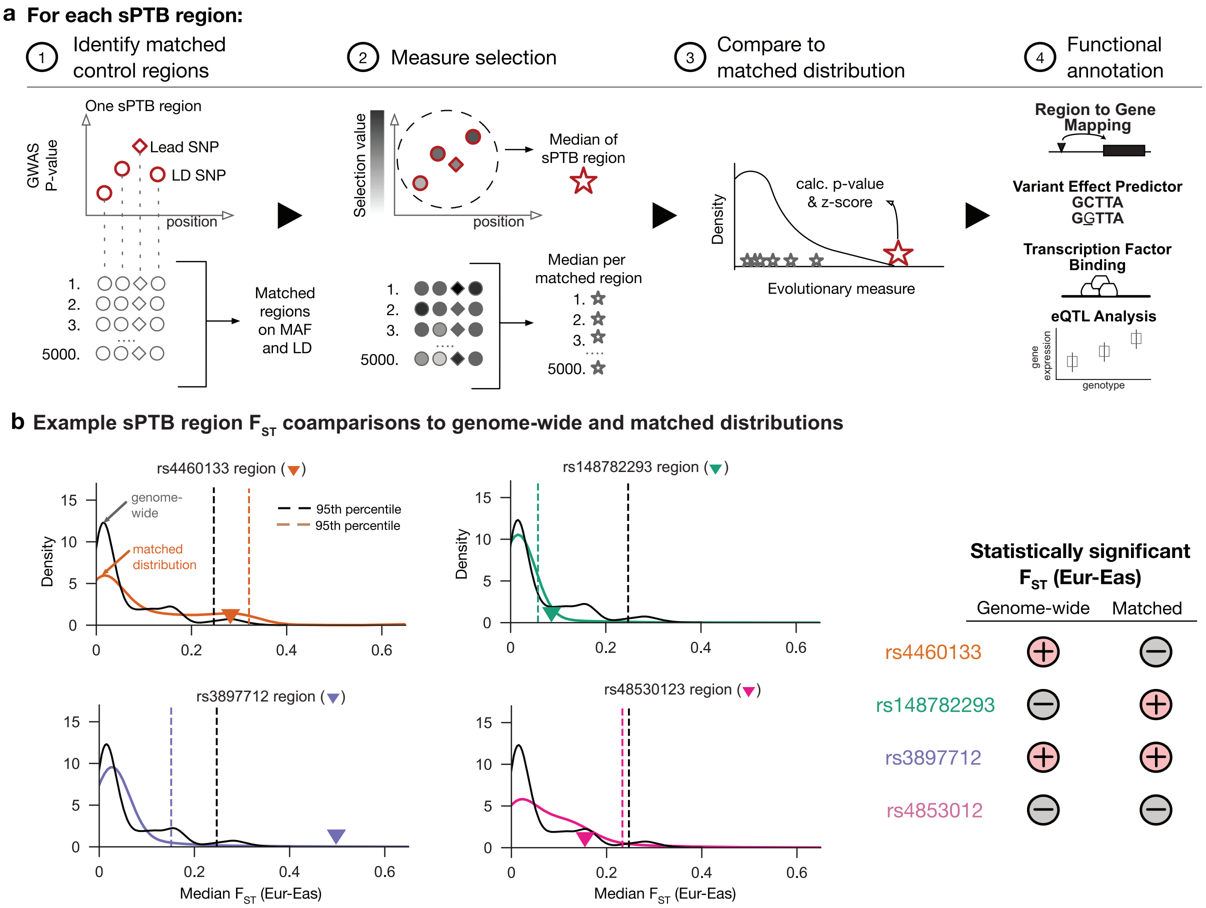


Figure 3.1: Accounting for minor allele frequency and linkage disequilibrium of genomic regions to identify loci that have experienced diverse evolutionary forces. (a) We compared evolutionary measures for each sPTB-associated genomic region (n=215) to ~5000 MAF and LD matched control regions. The sPTB-associated genomic regions each consisted of a lead variant (p<10E-4 association with sPTB) and variants in high LD (r^2>0.9) with the lead variant. Each control region has an equal number of variants as the corresponding sPTB-associated genomic region and is matched for MAF and LD ('Identify matched control regions'). We next obtained the values of an evolutionary measure for the variants included in the sPTB-associated regions and all control regions ('Measure selection'). The median value of the evolutionary measure across variants in the sPTB-associated region and all control regions was used to derive an empirical p-value and z-score ('Compare to matched distribution'). We repeated these steps for each sPTB-associated region and evolutionary measure and then functionally annotated sPTB-associated regions with absolute z-scores ≥ 1.5 ('Functional annotation'). (b) Representative examples for four sPTB-associated regions highlight differences in the distribution of genome-wide and matched control regions for an evolutionary measure (F_{ST} between Europeans and East Asians). The black and colored distributions correspond to genome-wide and matched distributions, respectively. The colored triangle denotes the median F_{ST} (Eur-Eas) for the sPTB-associated region. The dashed vertical lines mark the 95th percentile of the genome-wide (black) and matched (colored) distributions. If this value is greater than the 95th percentile, then it is considered significant (+); if it is lower than the 95th percentile it is considered not-significant (-). The four examples illustrate the importance of the choice of background in evaluating significance of evolutionary metrics (table to the right).

3.3 sPTB regions have been shaped by diverse modes of selection

To gain insight into the modes of selection that have acted on sPTB-associated genomic regions, we focused on genomic regions with extreme evolutionary signatures by selecting the 120 sPTB-associated regions with at least one extreme z-score ($z \geq +/- 1.5$) for an evolutionary metric (Figure 3.2) for further analysis. The extreme z-score for each of these 120 sPTB-associated regions suggests that the evolutionary force of interest has likely influenced this region when compared to the matched control regions. Notably, each evolutionary measure had at least one genomic region with an extreme observation ($p<0.05$). Hierarchical clustering of the 120 regions revealed 12 clusters of regions with similar evolutionary patterns. We manually combined the 12 clusters based on their dominant evolutionary signatures into five major groups with the following general evolutionary patterns (Figure 3.2): conservation/negative selection (group A: clusters A1-4), excess population differentiation/local adaptation (group B: clusters B1-2), positive selection (group C: cluster C1), long-term balanced polymorphism/balancing selection (group D: clusters D1-2), and other diverse evolutionary signatures (group E: clusters E1-4).

Previous literature on complex genetic traits[110–112] and pregnancy disorders[90,102,104,113] supports the finding that multiple modes of selection have acted on sPTB-associated genomic regions. Unlike many of these previous studies that tested only a single mode of selection, our approach tested multiple modes of selection. Of the 215 genomic regions we tested, 9% had evidence of conservation, 5% had evidence of excess population differentiation, 4% had evidence of accelerated evolution, 4% had evidence of long-term balanced polymorphisms, and 34% had evidence of other combinations. From these data we infer that negative selection, local adaptation, positive selection, and balancing selection have all acted on genomic regions associated with sPTB, highlighting the mosaic nature of the evolutionary forces that have shaped this trait.

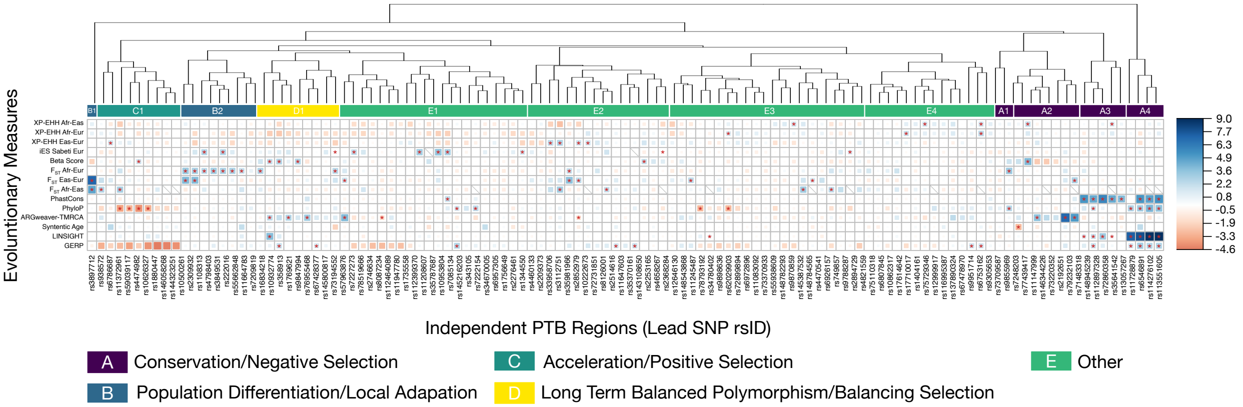


Figure 3.2: sPTB-associated genomic regions have experienced diverse evolutionary forces. We tested sPTB-associated genomic regions (x-axis) for diverse types of selection (y-axis), including FST (population differentiation), XP-EHH (positive selection), Beta Score (balancing selection), allele age (time to most recent common ancestor, TMRCA, from ARGweaver), alignment block age, phyloP (positive/negative selection), GERP, LINSIGHT, and PhastCons (negative selection) (Table 1, Figure 1). The relative strength (size of colored square) and direction (color) of each evolutionary measure for each sPTB-associated region is summarized as a z-score calculated from that region's matched background distribution. Only regions with |z| ≥ 1.5 for at least one evolutionary measure before clustering are shown. Statistical significance was assessed by comparing the median value of the evolutionary measure to the matched background distribution to derive an empirical p-value (*p>0.05). Hierarchical clustering of sPTB-associated genomic regions on their z-scores identifies distinct groups or clusters associated with different types of evolutionary forces. Specifically, we interpret regions that exhibit higher than expected values for PhastCons, PhyloP, LINSIGHT, and GERP to have experienced conservation and negative selection (Group A); regions that exhibit higher than expected pairwise FST values to have experienced population differentiation/local adaptation (Group B); regions that exhibit lower than expected values for PhyloP to have experienced acceleration/positive selection (Group C); and regions that exhibit higher than expected Beta Score and older allele ages (TMRCA) to have experienced balancing selection (Group D). The remaining regions exhibit a variety of signatures that are not consistent with a single evolutionary mode (Group E).

In addition to differences in evolutionary measures, variants in these groups also exhibited differences in their functional effects, likelihood of influencing transcriptional regulation, frequency distribution between populations, and effects on tissue-specific gene expression (Figure 3.3). Given that our starting dataset was identified using GWAS, we do not know how these loci influence sPTB. Using the current literature to inform our evolutionary analyses allows us to make hypotheses about links between these genomic regions and sPTB. In the next section, we describe each group and give examples of their members and their potential connection to preterm birth and pregnancy.

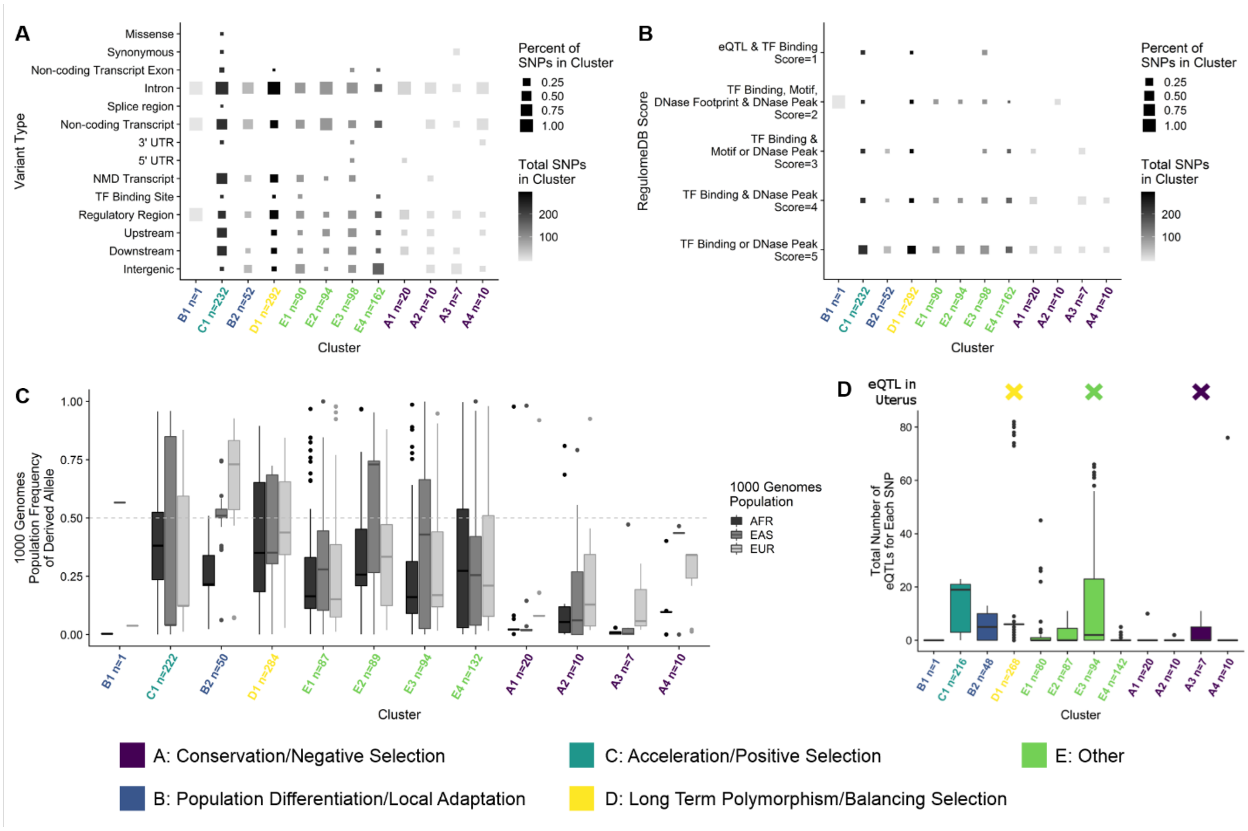


Figure 3.3: Clusters of preterm birth regions that have experienced different types of selection vary in their molecular characteristics and functions. Clusters are ordered as they appear in the z-score heatmap (Figure 2) and colored by their major type of selection: Group A: Conservation and negative selection (Purple), Group B: Population differentiation/local adaptation (Blue), Group C: Acceleration and positive selection (Teal), Group D: Long term polymorphism/balancing selection (Teal), and Other (Green). A. The proportions of different types of variants (e.g., intronic, intergenic, etc.) within each cluster (x-axis) based on the Variant Effect Predictor (VEP) analysis. Furthermore, cluster C1 exhibits the widest variety of variant types and is the only cluster that contains missense variants. Most variants across most clusters are located in introns. B. The proportion of each RegulomeDB score (y-axis) within each cluster (x-axis). Most notably, preterm birth regions in three clusters (B1, A5, and D4) have variants that are likely to affect transcription factor binding and linked to expression of a gene target (Score=1). Almost all clusters contain some variants that are likely to affect transcription factor binding (Score=2). C. The derived allele frequency (y-axis) for all variants in each cluster (x-axis) for the African (AFR), East Asian (EAS), and European (EUR) populations. Population frequency of the derived allele varies within populations from 0 to fixation. D. The total number of eQTLs (y-axis) obtained from GTEx for all variants within each cluster (x-axis) All clusters but one (C2 with only one variant) have at least one variant that is associated with the expression of one or more genes in one or more tissues. Clusters A1, A5, and D4 also have one or more variants associated with expression in the uterus.

3.4 Functional and evolutionary characteristics of sPTB associated regions

3.4.1 Group A: Sequence conservation/negative selection

Group A contained 19 genomic regions and 47 variants with higher than expected values for evolutionary measures of sequence conservation and alignment block age (Figure 3.2; Figure 3.3B), suggesting that these genomic regions evolved under negative selection. The action of negative selection is consistent with previous studies of sPTB associated genes[70]. The majority of variants are intronic (37/47: 79%) but a considerable fraction is intergenic (8/47: 17%; Figure 3.3).

In this group, the sPTB-associated variant (rs6546891, OR: 1.13; adjusted p-value: 5.4x10-5)[47] is located in the 3'UTR of the gene *TET3*. The risk allele (G) originated in the human lineage and is at lowest frequency in the European population. Additionally, this variant is an eQTL for 76 gene/tissue pairs and associated with gene expression in reproductive tissues, such as expression of *NAT8* in the testis. In mice, *TET3* had been shown to affect epigenetic reprogramming, neonatal growth, and fecundity[114,115]. In humans, *TET3* expression was detected in the villus cytotrophoblast cells in the first trimester as well as in maternal decidua of placentas[116]. *TET3* expression has also been detected in pathological placentas[117], and has also been linked to neurodevelopment disorders and preterm birth[118]. Similarly, *NAT8* is involved in epigenetic changes during pregnancy[119].

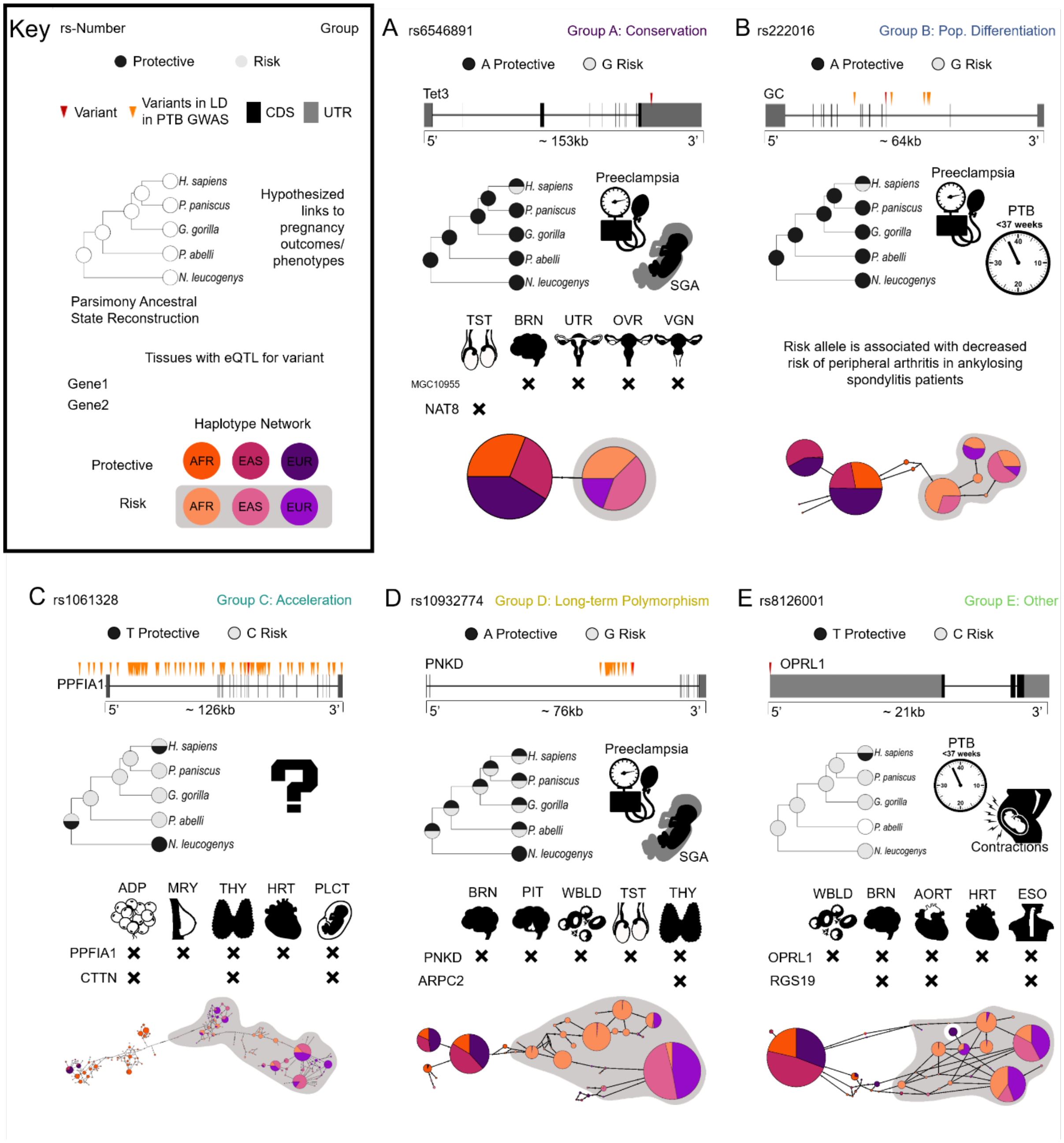


Figure 3.4: Functional and evolutionary characterizations of sPTB-associated genomic regions. For each variant we report the protective and risk alleles from the sPTB GWAS[47] the location relative to the nearest gene and linked variants; the alleles at this variant across the great apes and the parsimony reconstruction of the ancestral allele(s); hypothesized links to pregnancy outcomes or phenotypes; selected significant GTEx hits; and human haplotype(s) containing each variant in a haplotype map. a Group A (conservation): Human-specific risk allele of rs6546894 is located in the 3' UTR of TET3. TET3 expression is elevated in preeclamptic and small for gestational age (SGA) placentas[118]. rs6546894 is also associated with expression of MGC10955 and NAT8 in the testis (TST), brain (BRN), uterus (UTR), ovaries (OVR), and vagina (VGN). b Group B (population differentiation): rs222016, an intronic variant in gene GC, has a human-specific protective allele. GC is associated with sPTB [116] . c Group C (acceleration): rs1061328 is located in a PPFIA1 intron and is in LD with 156 variants. The protective allele is human-specific. This variant is associated with changes in expression of PPFIA1 and CTTN in adipose 63,108 cells (ADP), mammary tissue (MRY), the thyroid (THY), and heart (HRT). CTTN is

expressed the placenta [116,120]. d Group D (long-term polymorphism): rs10932774 is located in a PNKD intron and is in LD with 27 variants. Alleles of the variant are found throughout the great apes. PNKD is upregulated in severely eclamptic placentas[121] and ARPC2 has been associated with SGA[122] . Expression changes associated with this variant include PNKD and ARPC2 in the brain, pituitary gland (PIT), whole blood (WBLD), testis, and thyroid. e Group E (other): rs8126001 is located in the 5' UTR of OPRL1 and has a human-specific protective allele. The protein product of the ORPL1 gene is the nociceptin receptor, which is linked to contractions and the presence of [123,124] nociception in preterm uterus samples . This variant is associated with expression of OPRL1 and RGS19 in whole blood, the brain, aorta (AORT), heart, and esophagus (ESO).

3.4.2 Group B: Population differentiation/local adaptation

Group B (clusters B1 and B2) contained variants with a higher than expected differentiation (F_{ST}) between pairs of human populations (Figure 3.2). There were 10 sPTB-associated genomic regions in this group, which contain 53 variants. The majority of variants are an eQTL in at least one tissue (29/52; Figure 3.3D). The derived allele frequency in cluster B1 is high in East Asian populations and very low in African and European populations (Figure 3.3C). We found that 3 of the 10 lead variants have higher risk allele frequencies in African compared to European or East Asian populations. This is noteworthy because the rate of preterm birth is twice as high among black women compared to white women in the United States[6]. These three variants are associated with expression levels of the genes *SLC33A1*, *LOC645355*, and *GC*, respectively.

The six variants (labeled by the lead variant rs22016), within the sPTB-associated region near *GC*, Vitamin D Binding Protein, are of particular interest. The ancestral allele (G) of rs22201is found at higher frequency in African populations and is associated with increased risk of sPTB (European cohort, OR: 1.15; adjusted p-value 3.58x10-5; Figure 3.4B)[47]. This variant has been associated with vitamin D levels and several other disorders[125,126]. There is evidence that vitamin D levels prior to delivery are associated with sPTB[127], that levels of GC in cervico-vaginal fluid may help predict sPTB[116,128], and that vitamin D deficiency may contribute to racial disparities in birth outcomes. For example, vitamin D deficiency is a potential risk factor for preeclampsia among Hispanic and African American women[129]. The population-specific differentiation associated with variant rs222016 is consistent with the differential evolution of the vitamin D system between populations, likely in response to different environments and associated changes in skin pigmentation[130]. Our results add to the evolutionary context of the link between vitamin D and pregnancy outcomes[131] and suggest a role for variation in the gene *GC* in the ethnic disparities in pregnancy outcomes.

3.4.3 Group C: Accelerated substitution rates/positive selection

Variants in cluster C1 (group C) had lower than expected values of PhyloP. This group contained nine sPTB-associated genomic regions and 232 variants. The large number of linked variants is consistent with the accumulation of polymorphisms in regions undergoing positive selection. The derived alleles in this group show no obvious pattern in allele frequency between populations (Figure 3.3C). While most variants are intronic (218/232), there are missense variants in the genes Protein Tyrosine Phosphatase Receptor Type F Polypeptide Interacting Protein Alpha 1 (*PPFIA1)* and Plakophilin 1 (*PKP1;* Figure 3.3A). Additionally, 16 variants are likely to affect transcription factor binding (regulomeDB score of 1 or 2; Figure 3.3B). Consistent with this finding, 167/216 variants tested in GTEx are associated with expression of at least one gene in one tissue (Figure 3.3C).

The lead variant associated with *PPFIA1* (rs1061328) is linked to an additional 156 variants, which are associated with the expression of a total of 2,844 tissue/gene combinations. Two of these genes are cortactin (*CTTN*) and *PPFIA1,* which are both involved in cell adhesion and migration —critical processes in the development of the placenta and implantation[132,133]. Members of the PPFIA1 liprin family have been linked to maternal-fetal signaling during placental development[134,135], whereas *CTTN* is expressed in the decidual cells and spiral arterioles and localizes to the trophoblast cells during early pregnancy, suggesting a role for *CTTN* in cytoskeletal remodeling of the maternal-fetal interface[136]. There is also is evidence that decreased adherence of maternal and fetal membrane layers is involved in parturition[137]. Accelerated evolution has previously been detected in the birth timing-associated genes *FSHR*[138] and *PLA2G4C*. It has been hypothesized that human and/or primate-specific adaptations, such as bipedalism, have resulted in the accelerated evolution of birth-timing phenotypes along these lineages[139]. Accelerated evolution has also been implicated in other complex disorders—especially those like schizophrenia[140] and autism which affect the brain, another organ that is thought to have undergone adaptive evolution in the human lineage.

3.4.4 Group D: Balanced polymorphism/balancing selection

Variants in Group D generally had higher than expected values of beta score or an older than expected allele age, consistent with evolutionary signatures of balancing selection (Figure 3.2). There were nine genomic regions in group D; three had a significantly higher than expected beta scores ($p<0.05$), three have a significantly older than expected TMRCA values ($p<0.05$), and three have older TMRCA values but are not significant. The derived alleles have an average allele frequency across all populations of 0.44 (Figure 3.3C). GTEx analysis supports a regulatory role for many of these variants—266 of 271 variants are an eQTL in at least one tissue (Figure 3.3D).

The genes associated with the variant rs10932774 (OR: 1.11, adjusted p-value 8.85x10^{-547}; *PNKD* and *ARPC2*) show long-term evolutionary conservation consistent with a signature of balancing selection and prior research suggests links to pregnancy through a variety of mechanisms. For example, *PNKD* is up-regulated in severely preeclamptic placentas[121] and in PNKD patients pregnancy is associated with changes in the frequency or severity of PNKD attacks. Similarly, the Arp2/3 complex is important for early embryo development and preimplantation in pigs and mice[141,142], and *ARPC2* transcripts are subject to RNA editing in placentas associated with intrauterine growth restriction/small for gestational age[142]. The identification of balancing selection acting on sPTB-associated genomic regions is consistent with the critical role of the immune system, which often experiences balancing selection[143,144], in establishing and maintaining pregnancy. Overall, *PNKD* and *ARPC2* show long-term evolutionary conservation consistent with a signature of balancing selection and prior research suggests links to pregnancy through a variety of mechanisms. The identification of balancing selection acting on sPTB-associated genomic regions is consistent with the critical role of the immune system, which often experiences balancing selection[143,145], in establishing and maintaining pregnancy.

3.4.5 Group E: Varied evolutionary signatures

The final group, group E, contained the remaining genomic regions in clusters E1, E2, E3 and E4 and was associated with a broad range of evolutionary signatures (Figure 3.2). At least one variant in group E had a significant p-value for every evolutionary measure (except for alignment block age), 39 / 73 lead variants had a significant p-value ($p<0.05$) for either genomic evolutionary rate profiling (GERP) or cross-population extended haplotype homozygosity XP-EHH, and 23 / 33 genomic regions had high z-scores ($|z|>1.5$) for population-specific iHS. The high frequency of genomic regions with significant XP-EHH or population-specific iHS values suggests that population-specific evolutionary forces may be at play in this group and that that pregnancy phenotypes in individual populations may have experienced different mosaics of evolutionary forces, consistent with previous work that sPTB risk varies with genomic background[146,147]. Finally, there are 143 variants identified as eQTLs, including 16 expression changes for genes in the uterus (all associated with the variant rs12646130; Figure 3.4D).

Interestingly, this group contained variants associated with the *EEFSEC*, *ADCY5*, and *WNT4* genes, which have been previously associated with gestational duration or preterm birth[148]. The group E variant rs8126001 (effect: 0.896; adjusted p-value 4.04x10^{-5})[47] is located in the 5' UTR of the opioid related nociception receptor 1 or nociception opioid receptor (*OPRL1* or *NOP-R*) gene which may be involved in myometrial contractions during delivery[124]. This variant has signatures of positive selection as detected by the integrated haplotype score (iHS) within the African population (Supplementary Data 2) and is associated with expression of *OPRL1* in multiple tissues (Figure 4E). *OPRL1* encodes a receptor for the endogenous peptide nociceptin (N/OFQ), which is derived from prenociceptin (PNOC). N/OFQ and PNOC are detected in human pregnant myometrial tissues[47] and *PNOC* mRNA levels are significantly higher in human preterm uterine samples and can elicit myometrial relaxation *in vitro*[123]. It is therefore likely that nociceptin and *OPRL1* are involved in the perception of pain during delivery and the initiation of delivery.

3.5 sPTB loci are enriched for diverse evolutionary signatures

Our analyses have so far focused on evaluating the evolutionary forces acting on individual sPTB-associated regions. To test whether the entire set of sPTB-associated regions is enriched for specific evolutionary signatures, we compared the set to the genome-wide background as well as to matched background sets.

To compare the number of sPTB loci with evidence for each evolutionary force to the rest of the genome, we computed each metric on 5,000 randomly selected regions and report the number of the 215 sPTB loci in the top 5th percentile for each evolutionary measure. If the evolutionary forces acting on the sPTB loci are similar to those on the genomic background, we would expect 5% (~11 / 215) to be in the top tail. Instead, out of 215 sPTB regions tested, 26 regions on average are in the top 5th percentile across all evolutionary measures (Figure 3.5a). To generate confidence intervals for these estimates, we repeated this analysis 1,000 times and found that variation is low (S.D. ≤ 1 region). This demonstrates that, compared to genome-wide distribution, sPTB loci are enriched for diverse signatures of selection (Figure 3.5a).

To compare the number of sPTB loci with extreme evolutionary signatures to the number expected by chance after matching on MAF and LD, we generated 215 random regions, compared them to their MAF and LD matched distributions, and repeated this process 1,000 times for each evolutionary measure. The number regions expected by chance varied from ~4 to 11 (Figure 3.5b). For most evolutionary measures, the observed number of sPTB regions with extreme values was within the expected range from the random regions. However, measures of sequence conservation (LINSIGHT, GERP, PhastCons) and substitution rate (PhyloP) had more regions that were significant than expected by chance (top 5th percentile of the empirical distribution). Thus, sPTB regions are enriched for these evolutionary signatures compared to LD and MAF matched expectation (Figure 3.5b).

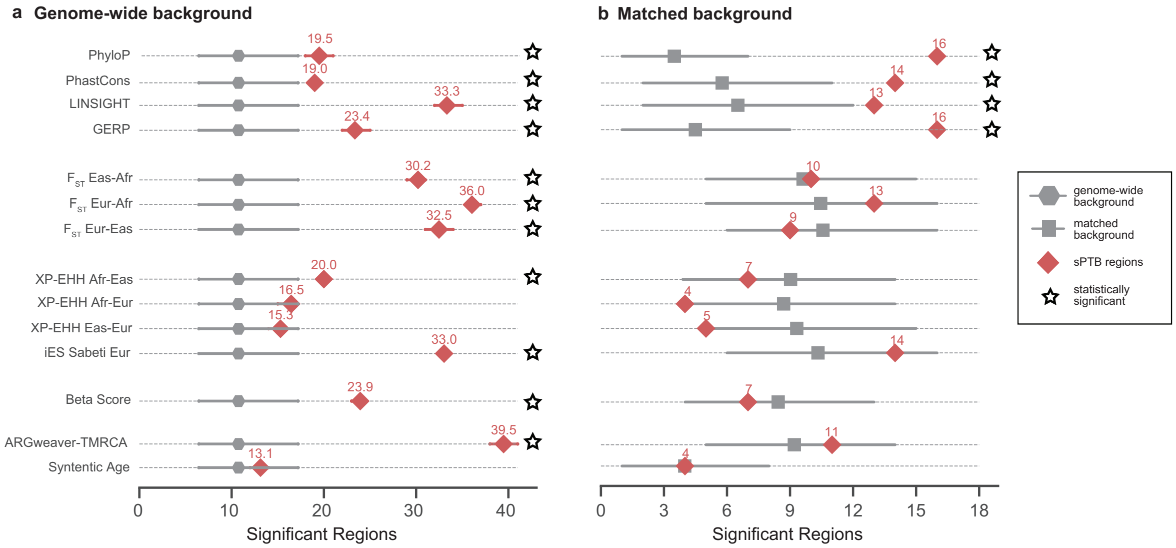


Figure 3.5: sPTB regions are enriched for diverse evolutionary measures compared to genome-wide distributions and for measures of sequence conservation when accounting for MAF and LD. (a) sPTB regions (red) are enriched for significant evidence of nearly all evolutionary measures (black stars) compared to the expectation from the genome-wide background (gray). For each evolutionary measure (y-axis), we evaluated the number of sPTB regions with statistically significant values ($p<0.05$) compared to the genome-wide distribution of the metric based on 5,000 randomly selected regions over 1,000 iterations. The mean number of significant regions (x-axis) is denoted by the red diamond with the 5th and 95th percentiles flanking. The expected number of significant regions by chance was computed from the binomial distribution (gray hexagons with 95% confidence intervals). (b) Accounting for MAF and LD revealed enrichment for evolutionary measures of sequence conservation (PhyloP, PhastCons, LINSIGHT, GERP) among the sPTB-associated genomic regions. In contrast to (a), the number of significant regions (among the 215 sPTB-associated regions) was determined based on 5,000 MAF- and LD-matched sets. Similarly, the expected distribution (gray boxes) was determined using 1,000 randomly selected region sets of the same size as the sPTB regions with matching MAF and LD values.

3.6 Discussion

In this study, we developed an approach to test for signatures of diverse evolutionary forces that explicitly accounts for MAF and LD in trait-associated genomic regions. Our approach has several advantages. First, for each genomic region associated with a trait, our approach evaluates the region's significance against a distribution of matched control genomic regions (rather than against the distribution of all trait-associated region or against a genome wide background, which is typical of outlier-based methods), increasing its sensitivity and specificity. Second, comparing evolutionary measures against a null distribution that accounts for MAF and LD further increases the sensitivity with which we can infer the action of evolutionary forces on sets of genomic regions that differ in their genome architectures. Third, because the lead SNPs assayed in a GWAS are often not causal variants, by testing both the lead SNPs and those in LD when evaluating a genomic region for evolutionary signatures, we are able to better represent the trait-associated evolutionary signatures compared to other methods that evaluate only the lead variant[70] or all variants, including those not associated with the trait, in a genomic window[149]. Fourth, our approach uses an empirical framework that leverages the strengths of diverse existing evolutionary measures and that can easily accommodate the additional of new evolutionary measures. Fifth, our approach tests whether evolutionary forces have acted (and to what extent) at two levels; at the level of each genomic region associated with a particular trait (e.g., is there evidence of balancing selection at a given region?), as well as at the level of the entire set of regions associated with the trait (e.g., is there enrichment for regions showing evidence of balancing selection for a given trait?). Finally, our approach can be applied to any genetically complex trait, not just in humans, but in any organism for which genome-wide association and sequencing data are available.

Although our method can robustly detect diverse evolutionary forces and be applied flexibly to individual genomic regions or entire sets of genomic regions, it also has certain technical limitations. The genomic regions evaluated for evolutionary signatures must be relatively small (r^2>0.9) in order to generate well-matched control regions on minor allele frequency and linkage disequilibrium. For regions with complex haplotype structures, this relatively small region may not tag the true effect-associated variant. Furthermore, since each genomic region has its own matched set of control regions, the computation burden increases with the number of trait-associated regions and the number of evolutionary measures. For each evolutionary measure, we must also be able to calculate its value for a large fraction of the control region variants. Although not all evolutionary measures can be incorporated into our approach, we demonstrate this approach on a large number of sPTB-associated regions across 11 evolutionary measures.

To illustrate our approach's utility and power, we applied it to examine the evolutionary forces that have acted on genomic regions associated with sPTB, a complex disorder of global health concern with a substantial heritability[150]. We find evidence of evolutionary conservation, excess population differentiation, accelerated evolution, and balanced polymorphisms in sPTB-associated genomic regions, suggesting that no single evolutionary force is responsible for shaping the genetic architecture of sPTB; rather, sPTB has been influenced by a diverse mosaic of evolutionary forces We hypothesize that the same is likely true of other complex human traits. While many studies have quantified the effect of selection on trait-associated regions[70,90,91], there are few tools available to concurrently evaluate multiple evolutionary forces as we have done here[107]. Deciphering the mosaic of evolutionary forces that have acted on human traits not only more accurately portrays the evolutionary history of the trait, but is also likely to reveal important functional insights and generate new biologically relevant hypotheses.

3.7 Methods

3.7.1 Deriving sPTB genomic regions from GWAS summary statistics

To evaluate evolutionary history of sPTB on distinct regions of the human genome, we identified genomic regions from the GWAS summary statistics. Using PLINK1.9b (pngu.mgh.harvard.edu/purcell/plink/)[85], the top 10,000 variants associated with sPTB from Zhang et. al. were clumped based on LD using default settings except requiring a p-value ≤ 10E-4 for lead variants and variants in LD with lead variants[104]. We used this liberal p-value threshold to increase the number of sPTB-associated variants evaluated. Although this will increase the number of false positive variants associated with sPTB, we anticipate that these false positive variants will not have statistically significant evolutionary signals using our approach to detect evolutionary forces. This is because the majority of the genome is neutrally evolving and our approach aims to detect deviation from this genomic background. Additionally, it is possible that the lead variant (variant with the lowest p-value) could tag the true variant associated with sPTB within an LD block. Therefore, we defined an independent sPTB-associated genomic region to include the lead and LD (r^2>0.9, p-value <= 10E-4) sPTB variants. This resulted in 215 independent lead variants within an sPTB-associated genomic region.

3.7.2 Creating matched control regions for sPTB-associated regions

We detected evolutionary signatures at genomic regions associated with sPTB by comparing them to matched control sets. Since many evolutionary measures are influenced by LD and allele frequencies and these also influence power in GWAS, we generated control regions matched for these attributes for observed sPTB-associated genomic regions. First, for each lead variant we identified 5,000 control variants matched on minor allele frequency (+/-5%), LD (r^2>0.9, +/-10% number of LD buddies), gene density (+/- 500%) and distance to nearest gene (+/-500%) using SNPSNAP[151], which derives controls variants from a quality controlled phase 3 100 Genomes (1KG) data, with default settings for all other parameters and the hg19/GRCh37 genome assembly. For each control variant, we randomly selected an equal number of variants in LD (r^2>0.9) as sPTB-associated variants in LD with the corresponding lead variant. If no matching control variant existed, we relaxed the LD required to r^2=0.6. If still no match was found, we treated this as a missing value. For all LD calculations, control variants and downstream evolutionary measure analyses, the European super-population from phase 3 1KG[152] was used after removing duplicate variants.

3.7.3 Evolutionary measures

To characterize the evolutionary dynamics at each sPTB-associated region, we evaluated diverse evolutionary measures for diverse modes of selection and allele history across each sPTB-associated genomic region. Evolutionary measures were either calculated or pre-calculated values were downloaded for all control and sPTB-associated variants. Pairwise Weir and Cockerham's F_{ST} values between European, East Asian, and African super populations from 1KG were calculated using VCFTools (v0.1.14)[152,153]. Evolutionary measures of positive selection, integrated haplotype score (iHS), XP-EHH, and integrated site-specific EHH (iES), were calculated from the 1KG data using rehh 2.0[152,154]. Beta score, a measure of balancing selection, was calculated using BetaScan software [152,155]. Alignment block age was calculated using 100-way multiple sequence alignment[156] to measure the age of alignment blocks defined by the oldest most recent common ancestor. The remaining measures were downloaded from publicly available sources: phyloP and phastCons 100 way alignment from UCSC genome browser[157]; LINSIGHT[158]; GERP [159,160]; and allele age (time to most common recent ancestor from ARGWEAVER)[89]. Due to missing values, the exact number of control regions varied by sPTB-associated region and evolutionary measure. We first marked any control set that did not match at least 90% of the required variants for a given sPTB-associated region, then any sPTB-associated region with ≥ 60% marked control regions were removed for that specific evolutionary measure. iHS was not included in Figure 3.2 because of large amounts of missing data for up to 50% of genomic regions evaluated.

3.7.4 Detecting significant differences in evolutionary measures

For each sPTB-associated genomic region for a specific evolutionary measure, we took the median value of the evolutionary measure across all variants in LD in the region and compared it to the distribution of median values from the corresponding MAF- and LD-matched control regions described above. Statistical significance for each sPTB-associated region was evaluated by comparing the median value of the evolutionary measure to the distribution of median values of the control regions. To obtain the p-value, we calculated the number of control regions with a median value that are equal to or greater the median value for the preterm birth region. Since allele age (time to most recent common ancestor (TMRCA) from ARGweaver), PhyloP, and alignment block age are bi-directional measures, we calculated two-tailed p-values; all other evolutionary measures used one-tailed p-values. To compare evolutionary measures whose scales differ substantially, we calculated a z-score for each region per measure. These z-scores were hierarchically clustered across all regions and measures. Clusters were defined by a branch length cutoff of seven. These clusters were then grouped and annotated by the dominant evolutionary measure through manual inspection to highlight the main evolutionary trend(s).

3.7.5 Annotation of variants in sPTB-associated regions

To understand functional differences between groups and genomic regions we collected annotations for variants in sPTB-associated regions from publicly available databases. Evidence for regulatory function for individual variants was obtained from RegulomeDB v1.1 (accessed 1/11/19)[161]. From this we extracted the following information: total promotor histone marks, total enhancer histone marks, total DNase 1 sensitivity, total predicted proteins bound, total predicted motifs changed, and regulomeDB score. Variants were identified as expression quantitative trait loci (eQTLs) using the Genotype-Tissue Expression (GTEx) project data (dbGaP Accession phs000424.v7.p2 accessed 1/15/19). Variants were mapped to GTEx annotations based on RefSNP (rs) number and then the GTEx annotations were used to obtain eQTL information. For each locus, we obtained the tissues in which the locus was an eQTL, the genes for which the locus affected expression (in any tissue), and the total number times the locus was identified as an eQTL. Functional variant effects were annotated with the Ensembl Variant Effect Predictor (VEP; accessed 1/17/19) based on rs number[162]. Variant to gene associations were also assessed using GREAT[163]. Total evidence from all sources—nearest gene, GTEx,VEP, regulomeDB, GREAT—was used to identify gene-variant associations. Population-based allele frequencies were obtained from the 1KG phase3 data for the African (excluding related African individuals), East Asian, and European populations[152].

To infer the history of the alleles at each locus across mammals, we created a mammalian alignment at each locus and inferred the ancestral states. That mammalian alignment was built using data from the sPTB GWAS[47] (risk variant identification), the UCSC Table Browser [156] (30 way mammalian alignment), the 1KG phase 3[152] data (human polymorphism data) and the Great Ape Genome project (great ape polymorphisms)[164] which reference different builds of the human genome. To access data constructed across multiple builds of the human genome, we used Ensembl biomart release 97[165] and the biomaRt R package[166,167] to obtain the position of variants in hg38, hg19, and hg18 based on rs number[168]. Alignments with more than one gap position were discarded due to uncertainty in the alignment. All variant data were checked to ensure that each dataset reported polymorphisms in reference to the same strand. Parsimony reconstruction was conducted along a phylogenetic tree generated from the TimeTree database[169]. Ancestral state reconstruction for each allele was conducted in R using parsimony estimation in the phangorn package[170]. Five character-states were used in the ancestral state reconstruction: one for each base and a fifth for gap. Haplotype blocks containing the variant of interest were identified using Plink (v1.9b_5.2) to create blocks from the 1KG phase3 data. Binary haplotypes were then generated for each of the three populations using the IMPUTE function of vcftools (v0.1.15.) Median joining networks[171] were created using PopART[172].

3.7.6 Enrichment of significant evolutionary measures

Considering all sPTB regions, we evaluated whether sPTB regions overall are enriched for each evolutionary measure compared to genome-wide and matched control distributions. First, for the genome-wide comparisons, we counted the number of sPTB regions in the top 5th percentile of genome-wide distribution generated from 5,000 random regions for a given evolutionary measure. We repeated this step 1,000 times and computed the mean number of regions in the top 5th percentile of each iteration. The null expectation and statistical significance were computed using the Binomial distribution with a 5% success rate over 215 trials. Second, since many evolutionary measures are dependent on allele frequency and linkage equilibrium, we also compared the number of significant regions (over all sPTB regions) for an evolutionary measure to LD- and MAF-matched distributions as described earlier (Figure 3.1, Methods). To generate the null expectation for the number of significant regions, we randomly generated regions equal to the number of sPTB regions (n=215) and compared them to their own matched distributions. We repeated this for 1,000 sets of 215 random regions to generate the null distribution of the number of regions in the top 5th percentile for each evolutionary measure when matching for MAF and LD.

CHAPTER IV

4 Machine learning prediction of preterm births using EHRs linked to genetic biobanks[2]

4.1 Introduction

Preterm birth, occurring before 37 weeks of completed gestation, affects approximately 10% of pregnancies globally[2,4,31] and is the leading cause of infant mortality worldwide[1,173]. The causes of preterm birth are multifactorial, since different biological pathways and environmental exposures can trigger premature labor[49]. Large epidemiological studies have identified many risk factors, including multiple gestations[4], cervical anatomic abnormalities[174], and maternal age[175]. Notably, even though a history of preterm birth [27] is one of the strongest risk factors, the recurrence rate remains low at < 30%[176,177]. Additionally, maternal race is associated with risk for preterm birth; Black women have twice the prevalence compared to white women[4,6]. Preterm births have a heterogenous clinical presentation and cluster based on maternal, fetal, or placental conditions[31]. These obstetric and systemic comorbidities (e.g. pre-existing diabetes, cardiovascular disease) can also increase risk for preterm birth[30,178].

Despite our understanding of numerous risk factors, there are no accurate methods to predict preterm birth. Some biomarkers associate with preterm birth, but their best performance is limited to a subset of all cases[179,180]. Recently, analysis of maternal cell-free RNA has emerged as a promising approach[181], but initial results were based on a small pregnancy cohort and require further validation. In silico classifiers based on demographic and clinical risk factors have the advantage of not requiring serology or invasive testing. However, even in large cohorts (>1 million individuals), demographic- and risk-factor-based models report limited discrimination (AUC=0.63-0.74)[182–185]. To date, we lack effective screening tools and preventative strategies for prematurity[186].

[2] This work is a published as a preprint[79].

EHRs are scalable, readily available, and cost-efficient for disease-risk modeling[187]. EHRs capture longitudinal data across a broad set of phenotypes with detailed temporal resolution. EHR data can be combined with socio-demographic factors and family medical history to comprehensively model disease risk[188–190]. EHRs are also increasingly being augmented by linking patient records to molecular data, such as DNA and laboratory test results[191]. Since preterm birth has a substantial heritable risk[47], combining rich phenotypes with genetic risk may lead to better prediction.

Machine learning models have shown promise for accurate risk stratification across a variety of clinical domains[77,192,193]. However, despite the rapid adoption of machine learning in translational research, a review of 107 risk prediction studies reported that most models used only few variables, did not consider longitudinal data, and rarely evaluated model performance across multiple sites[194]. Some medical domains have yet to incorporate machine learning methods. Pregnancy research is especially well poised to benefit from machine learning approaches[188]. Per standard of care during pregnancy, women are carefully monitored with frequent prenatal visits, medical imaging, and clinical laboratories tests. Compared to other clinical contexts, pregnancy and the corresponding clinical surveillance occur in a defined timeframe based on gestational length. Thus, EHRs are well-suited for modeling pregnancy complications, especially when combined with the well documented outcomes at the end of pregnancy.

In this study, we combine multiple sources of data from EHRs to predict preterm birth using machine learning. From Vanderbilt's EHR database (>3.2 million records) and linked genetic biobank (>100,000 individuals), we identified a large cohort of women (n=35,282) with documented deliveries. We trained models (gradient boosted decision trees) that combine demographic factors, clinical history, laboratory tests, and genetic risk with billing codes (ICD-9 and CPT) to predict preterm birth. We find models trained on only billing codes show potential for predicting preterm birth. Billing code based models outperform a similar model using only known clinical risk factors. Across a variety of clinical contexts, such as second or spontaneous preterm birth, our models maintain accuracy. By investigating the patterns learned by our models, we identify clusters with distinct preterm birth risk and comorbidity profiles. Finally, we demonstrate the generalizability of our billing-code-based models on an external, independent cohort from the University of California, San Francisco (UCSF, n=5,978).

Prediction models trained at Vanderbilt maintain high accuracy in the external cohort with only a modest drop in performance. Our findings provide a proof-of-concept that machine learning on rich phenotypes in EHRs show promise for portable, accurate, and non-invasive prediction of preterm birth. The strong predictive performance across clinical context and preterm birth subtypes argues that machine learning models have the potential to add value during the management of pregnancy; however, further work is needed before these models can be applied in clinical settings.

4.2 Results

4.2.1 Assembling pregnancy cohort and ascertaining delivery type from Vanderbilt EHRs

From the Vanderbilt EHR database (>3.2 million patients), we identified a 'delivery cohort' of 35,282 women with at least one delivery in the Vanderbilt hospital system (Figure 4.1A). In addition to ICD and CPT billing codes, we extracted demographic data, past medical histories, obstetric notes, clinical labs, and genome-wide genetic data for the delivery cohort. Because billing codes were the most prevalent data in this cohort (n=35,282), we quantified the pairwise overlap between billing codes and each other data type. The largest subset included women with billing codes paired with demographic data (n=33,570). The smallest subset was women with billing codes paired with genetic data (n=905; Figure 4.1C). The mean maternal age at the first delivery in the delivery cohort was 27.3 years (23.0–31.0 years, 25th and 75th percentiles, Figure 4.21A). The majority of women in the cohort self- or third-party reported as white (n=21,343), Black (n=6,178), or Hispanic (n=3,979). The estimated gestational age (EGA) distribution had a mean of 38.5 weeks (38.0 to 40.3 weeks, 25th to 75th percentile; Figure 4.1D). The rate of multiple gestations (e.g. twins, triplets) was (7.6%, n=1,353). Since multiple gestation pregnancies are more likely to deliver preterm, we developed prediction models using singleton pregnancies unless otherwise stated.

To determine the delivery date and type (preterm vs. not-preterm) at scale across our large cohort, we developed a phenotyping algorithm using delivery-specific billing codes and estimated gestational age at delivery. For women with multiple pregnancies, we only considered the earliest delivery. We find that delivery-specific billing codes that can be used to label preterm births have high concordance (PPV≥0.85, Recall ≥0.95) with EGA based delivery labels (Figure 4.1E). Our final algorithm combined billing codes and EGA when available (n=15,041, Figure 4.1C). To evaluate the accuracy of the ascertained delivery labels, a domain expert blinded to the delivery type reviewed clinical notes from 104 EHRs selected at random from the delivery cohort. The algorithm had high accuracy: precision (positive predictive value) of 96% and recall (sensitivity) of 96% using the chart reviewed label as the gold standard (Figure 4.1F).

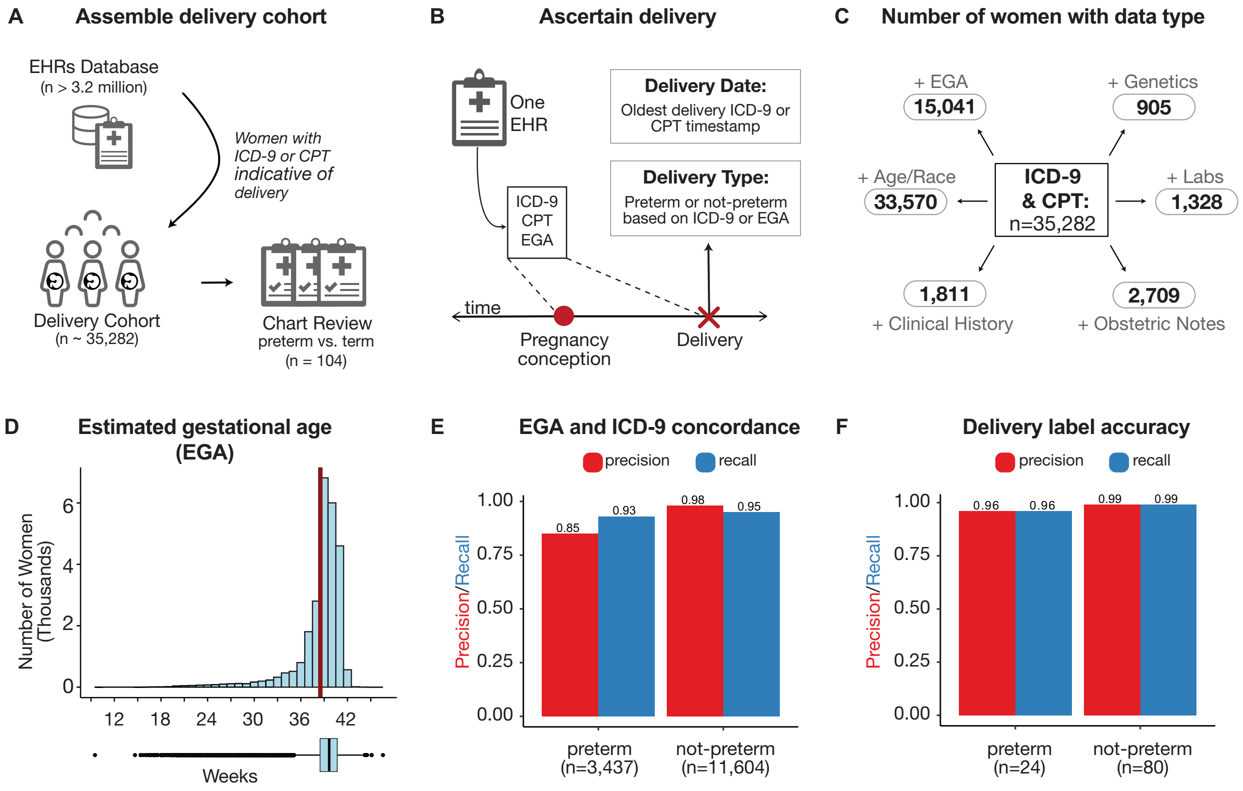


Figure 4.1: Schematic overview of the assembly of the delivery cohort from EHRs (EHRs). (A) Using billing codes, women with at least one delivery were extracted from the EHR database (n=35,282). (B) Delivery date and type were ascertained using ICD-9, CPT, and/or estimated gestational age (EGA) from each woman's EHR (Methods). From this cohort, 104 randomly selected EHRs were chart reviewed to validate the preterm birth label for the first recorded delivery. (C) Number of women in billing code cohort with estimated gestational age (+EGA), demographics (+Age, self- or third-party reported Race), clinical labs (+Labs), clinical obstetric notes (+Obstetric notes), patient clinical history (+Clinical History), and genetic data (+Genetics). (D) The EGA distribution at delivery (mean 38.5 weeks (red line); 38.0-40.3 weeks, 25th and 75th percentiles). Less than 0.015% (n=49) deliveries have EGA below 20 weeks. (E) The concordance between estimated gestational age (EGA) within three days of delivery and ICD-9 based delivery type for the 15,041 women with sufficient data for both. Precision and recall values were > 93% across labels except for preterm precision (85%). (F) Accuracy of delivery type phenotyping. The phenotyping algorithm was evaluated by chart review of 104 randomly selected women. The approach has high precision and recall for binary classification of 'preterm' or 'not-preterm'.

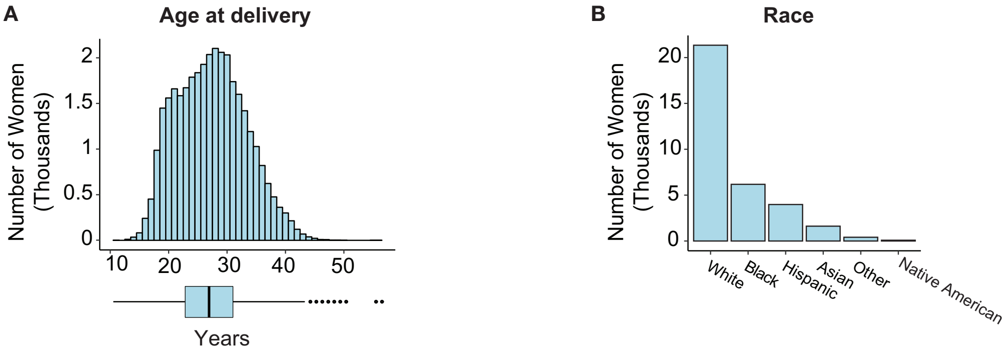
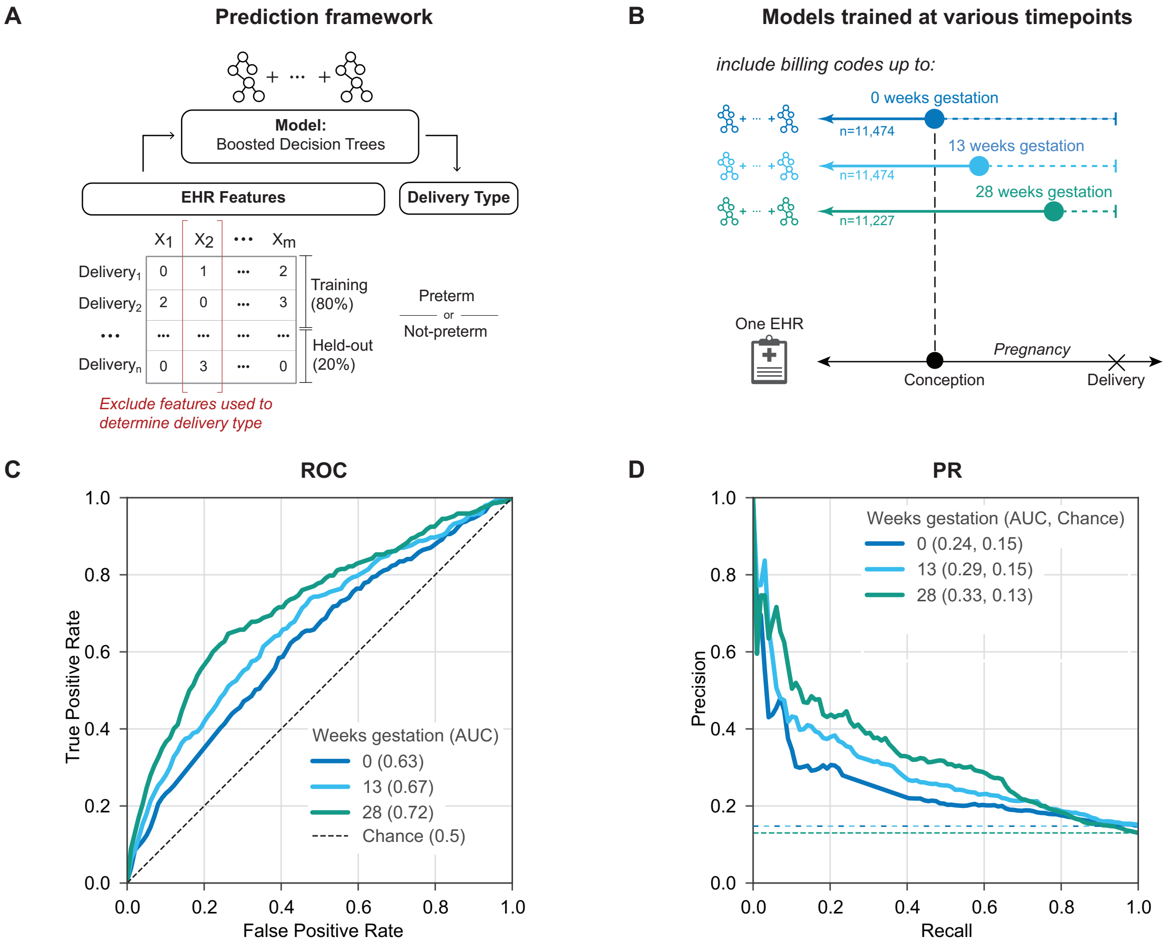


Figure 4.2: Distribution of maternal age at delivery and self- or third-party reported race. (A) The distribution of age at first delivery in EHR (mean 27.3 years; 23.0–31.0 years, 25th and 75th percentiles). (B) Counts of women by self- or third-party reported race (White: 21,343; Black: 6,178; Hispanic: 3,979; Asian: 1,617; Other: 409; Native American: 84).

4.2.2 Boosted decision trees using billing codes identify preterm deliveries

Using this richly phenotyped delivery cohort, we evaluated how well the entire clinical phenome, defined as billing codes (ICD-9 and CPT) before and after delivery, could identify preterm births. With counts of each billing code (excluding those used to ascertain delivery type), we trained gradient boosted decision trees[195] to classify each mother's first delivery as preterm or not-preterm. Boosted decisions trees are well-suited for EHR data because they require minimal transformation of the raw data, are robust to correlated features, and capture non-linear relationships[196]. Moreover, boosted decision trees have been successfully applied on a variety of clinical tasks[189,197,198].

In all evaluations, we held out 20% of the cohort as a test set and used the remaining 80% for training and validation (Figure 4.3A). Boosted decision tree models trained on ICD-9 and CPT codes accurately identified preterm births (singletons and multiple gestations) with PR-AUC=0.86 (chance=0.22) and ROC-AUC=0.95 (Figure 4.4C, B). While the combined ICD-9 and CPT based model achieved the best performance, models trained on either ICD-9 or CPT individually also performed well (PR-AUC ≥0.82; chance=0.22, ROC-AUC ≥0.93). All three models demonstrated good calibration with low Brier scores (≤0.092; Figure 4.4C). Thus, billing codes across an EHR show potential as a discriminatory feature for predicting preterm birth.



	X_1	X_2	···	X_m
$Delivery_1$	0	1	···	2
$Delivery_2$	2	0	···	3
···	···	···	···	···
$Delivery_n$	0	3	···	0

Figure 4.3: Machine learning classifiers accurately predict preterm birth using billing codes present before 28 weeks of gestation. (A) Machine learning framework for training and evaluating all models. We train models (boosted decision trees) on 80% of each cohort to predict the delivery as preterm or not-preterm. EHR features used to ascertain delivery type are excluded from training. Performance is reported on the held-out cohort consisting of 20% of deliveries using area under the ROC and precision-recall curves (ROC-AUC, PR-AUC). (B) We trained models using billing codes (ICD-9 and CPT) present before each of the following timepoints during pregnancy: 0, 13, and 28 weeks of gestation. These timepoints were selected to approximate gestational trimesters. Women who already delivered were excluded at each timepoint. To facilitate comparison across timepoints, we downsampled cohorts available so that the models were trained on a cohort with similar numbers of women (n=11,227 to 11,474). (C) The ROC-AUC increased from conception at 0 weeks (0.63, dark blue line) to 28 weeks of gestation (0.72, green line) compared to a chance (black dashed line) AUC of 0.5. (D) The model at 28 weeks of gestation achieved the highest PR-AUC (0.33). This is an underestimate of the possible performance; the accuracy improves further when all women with data available at 28 weeks are considered. Chance (dashed lines) represents the preterm birth prevalence in each cohort.

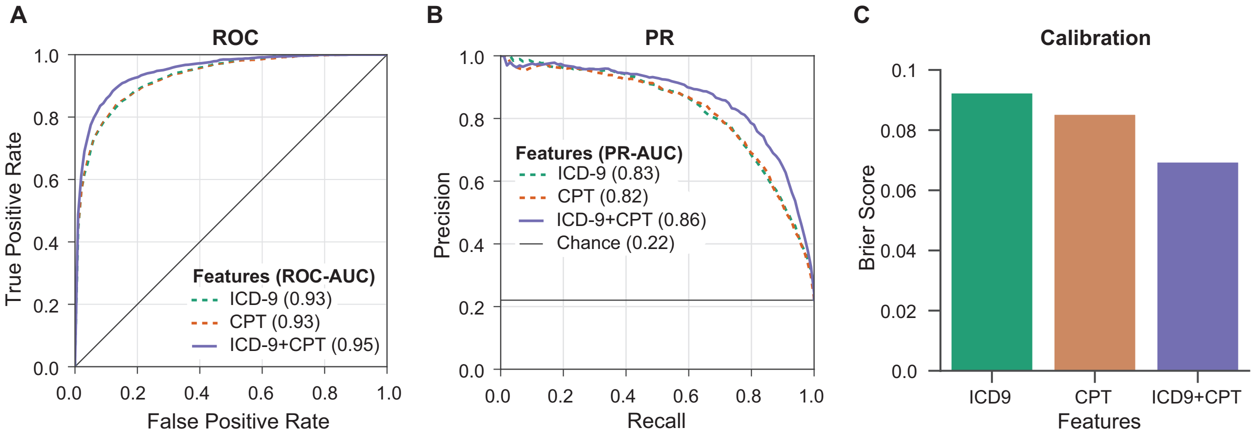


Figure 4.4: Boosted decision trees trained on EHR billing codes accurately identify preterm births. We trained and validated boosted decision trees on 80% of labeled pregnancies (preterm vs. non-preterm) from the EHR cohort (n=35,282, Fig. 1). We included both singletons and multiple gestations. We evaluated model performance on the held-out set using area under the ROC and precision-recall curves (ROC-AUC, PR-AUC) and Brier scores. EHR features used to ascertain delivery labels are excluded in training and evaluation of the models. (A,B) The boosted decision trees accurately classified deliveries by preterm birth status using only ICD-9 (green dashed line), only CPT (orange dashed line), and combined ICD-9 and CPT (solid purple) features present in a women's EHR, (ROC-AUC≥0.93, PR-AUC≥0.86). Combining ICD-9 and CPT codes achieved the best performance. (C) The low Brier scores (≤0.092) indicate that the models are well calibrated.

Accurate prediction of preterm birth at 28 weeks of gestation

To evaluate preterm birth prediction in a clinical context, we trained a boosted decision tree model (Figure 4.3A) on billing codes present before each of the following timepoints: 0, 13, and 28 weeks of gestation (Figure 4.3B). These timepoints were selected to approximately reflect pregnancy trimesters. We down-sampled to achieve comparable number of singleton deliveries across each timepoint (n=11,227 to 11,474) to mitigate sample size as a potential confounder while comparing performance. We only considered active pregnancies at each timepoint; for example, a delivery at 27 weeks would not be included in the 28-week model, since the outcome would already be known. The ROC-AUC increased from conception (0 weeks; 0.63) to the highest performance at 28 weeks (0.72; Figure 4.3C). The PR-AUC (Figure 4.3D), which accounts for preterm birth prevalence, is highest at 28 weeks (0.33, chance=0.13). However, as we show in the next section, this is an underestimate of the ability to predict preterm delivery at 28 weeks due to the down-sampling of the number of training examples. As expected, when we included multiple gestations, the model performed even better (PR-AUC=0.42 at 28 weeks, chance=0.14;Figure 4.5). Results were similar when models were trained using billing codes available before different timepoints from the date of delivery (Figure 4.6).

To test whether differences in contact with the health system between cases and controls were driving performance, we trained a classifier based on the total number of codes in an individual's EHR before delivery to predict preterm birth. This simple classifier failed to discriminate between delivery types with PR-AUC and ROC-AUC only slightly higher than chance (PR-AUC=0.19, chance=0.19; ROC-AUC=0.56, chance=0.5,Figure 4.7). Therefore, cumulative disease burden or the number of contacts alone are not informative for predicting preterm birth.

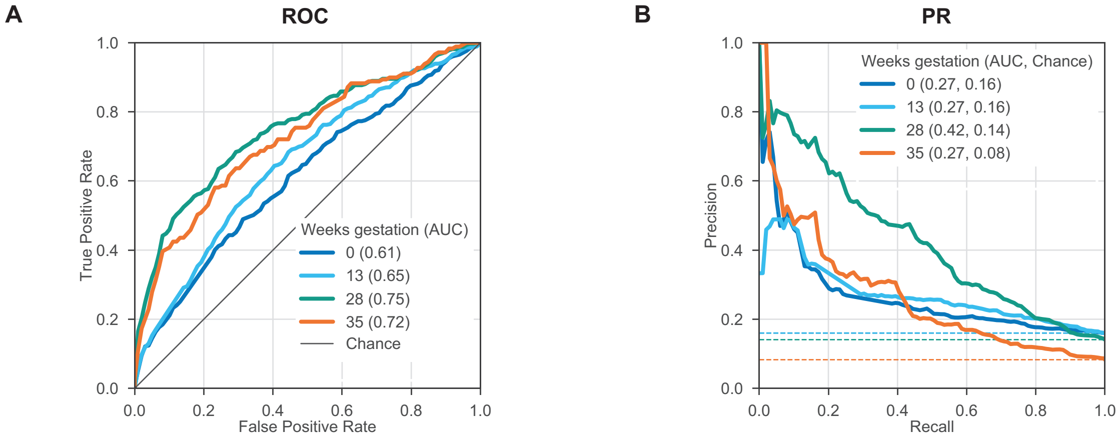


Figure 4.5: Machine learning can accurately identify preterm birth including singletons and multiple gestations. We trained models (boosted decision trees) on 80% of the corresponding cohort to predict the earliest delivery as preterm or not-preterm (Methods). In contrast to the models presented in the main text (Fig. 2), these included singleton and multiple gestations. Billing codes (ICD-9 and CPT) present before pregnancy (0, 13, 28, and 35 week of gestation) were used to train models. The same cohort of women (training + held-out) was used to train and evaluate across models but the sample size varied slightly (n = 11,843 to 10,799) since women who already delivered were excluded at each timepoint. (A) The ROC-AUC increased from conception at 0 weeks (0.61, dark blue line) to 35 weeks of gestation (0.72, orange line) compared to a chance (black line). (B) The model at 28 weeks of gestation achieved the highest PR-AUC (0.42). Chance (dashed lines) represents the preterm birth prevalence in each cohort.

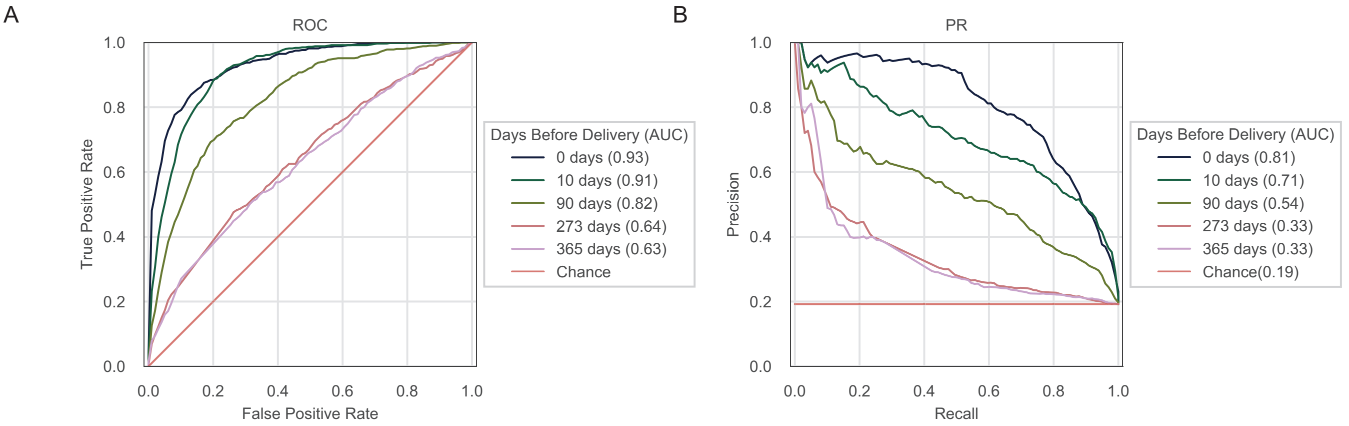


Figure 4.6: Preterm birth prediction increases at timepoints closer to the date of delivery at timepoints based on days before delivery. (A) ROC and (B) PR curves for preterm birth prediction using billing codes (ICD-9 & CPT) at different timepoints defined from the date of delivery in the Vanderbilt cohort. Both singletons and multiple gestations are included. Chance for PR-AUC represent random prediction equal to the population prevalence of preterm birth. Model performance improves as the prediction is made closer to delivery. All models are trained and evaluated on the same cohort of women (n=15,481) and the performance reported is on the held-out set (20% of cohort).

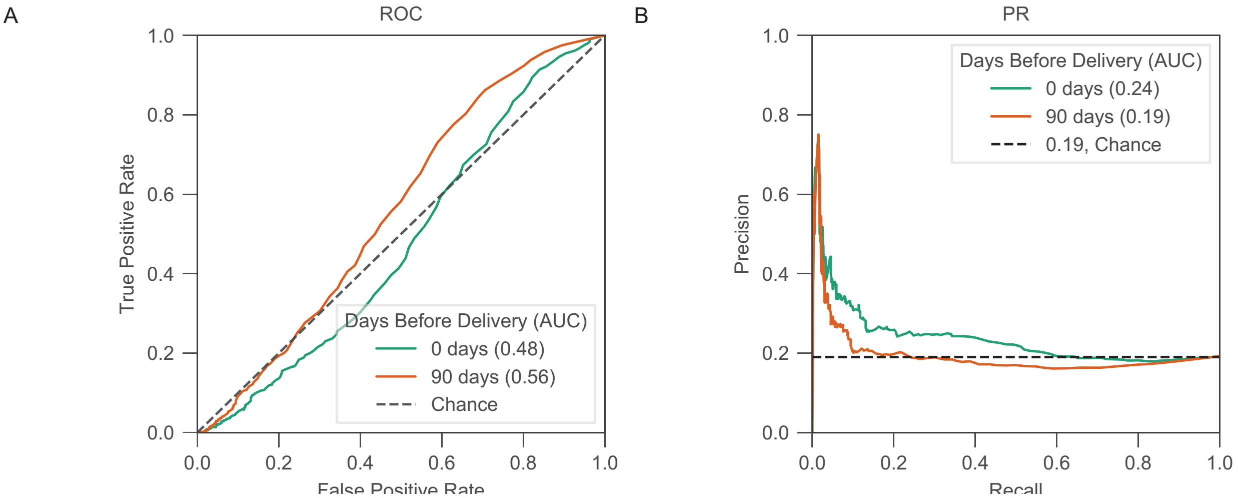


Figure 4.7: Preterm birth prediction is not driven by total number of billing codes.
To evaluate whether the amount of contact with the healthcare system was driving the performance of our machine learning classifiers, we assessed the discriminatory ability of the total number of billing codes (ICD-9 or CPT) in a woman's EHR to predict preterm birth. We include both singletons and multiple gestations. A simple classifier that used only the number of total billing codes preset at 0 days (green) and 90 days (orange) before the first delivery in her EHR, did not predict preterm birth well: (A) ROC-AUC = 0.56 and (B) PR-AUC = 0.19. The cohort consisted of the held-out set at the specified timepoints with 3,096 women.

4.2.3 Integrating other EHR features does not improve model performance

In addition to billing codes, EHRs capture aspects of an individual's health through different types of structured and unstructured data. We tested whether incorporating additional features from EHRs can improve preterm birth prediction. Models were evaluated using data available at 28 weeks of gestation; we selected this time point as a tradeoff between being sufficiently early for some potential interventions and late enough for sufficient data to be present to enable accurate predictions using billing codes. From the EHRs, we extracted sets of features including demographic variables (age, race), clinical keywords from obstetric notes, clinical lab tests ran during the pregnancy, and predicted genetic risk (polygenic risk score for preterm birth). To measure the performance gain for each feature set, we compared models trained using: the feature set only, billing codes only, and billing codes combined with the feature set (Figure 4.8A). Within each feature set, the same pregnancies comprised the training and held-out sets for the three models. However, the number of deliveries (training + held-out sets) varied widely across feature sets (n=462 to 20,342) due to the differing availability of each feature type.

Models using only demographic factors, clinical keywords, and genetic risk had ROC-AUC and PR-AUC similar to chance (Figure 4.8B). Clinical labs had moderate predictive power with ROC-AUC of 0.63 and PR-AUC of 0.24 (Figure 4.8B). Compared to models using only billing codes, adding additional feature sets did not substantially improve performance (Figure 4.8B). We note that some features sets, such as clinical labs and genetic risk, were evaluated on held-out sets with small numbers of deliveries (180 and 92, respectively). However, even after increasing the sample size by including women who may have features either before or after delivery, we did not observe a consistent gain in performance compared to models trained using only billing codes (Figure 4.9).

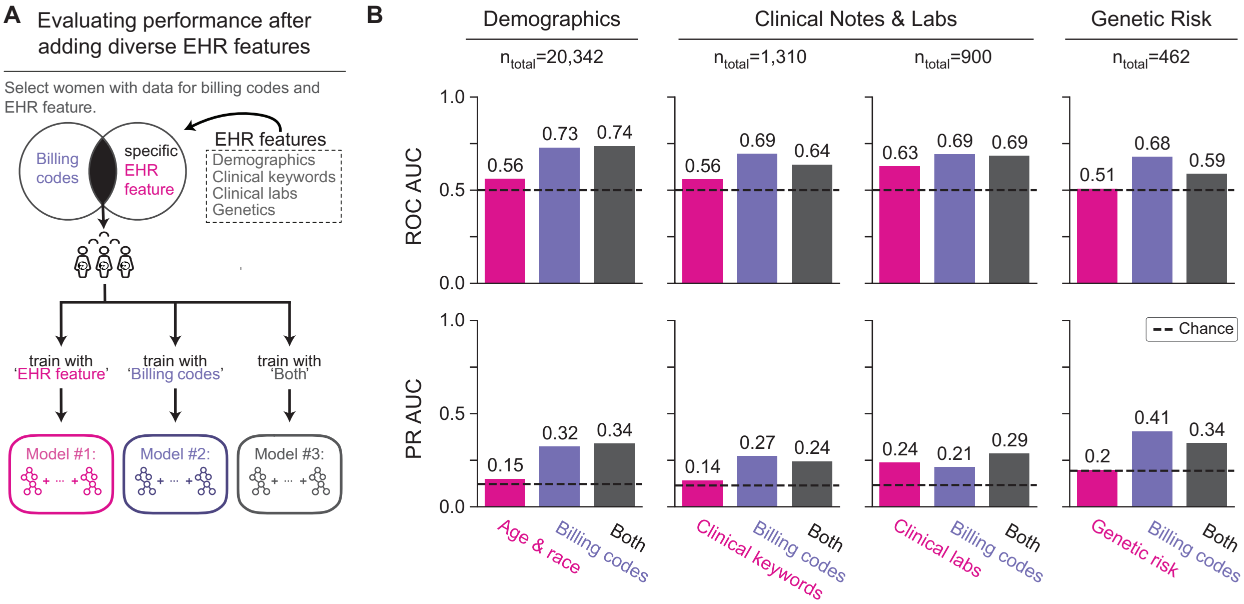


Figure 4.8: Combing demographic, clinical, and genetic features does not substantially improve preterm birth prediction compared to using only billing codes. (A) Framework for evaluating change in preterm birth prediction performance after incorporating diverse types of EHR features with billing codes (ICD-9 and CPT codes). We used only features and billing codes occurring before 28 weeks of gestation. EHR features are grouped by sets of demographic factors (age and race), clinical keywords (UMLS concept unique identifiers from obstetric notes), clinical labs, and genetic risk (polygenic risk score for preterm birth). We compared three models for each feature set: 1) using only the feature set being evaluated (pink), 2) using only billing codes ('Billing codes', purple), and 3) using the feature set combined with billing codes ('Both', gray). For each feature set, we considered the subset of women who had at least one recorded value for the EHR feature and billing codes. All three models for a given EHR feature set considered the same pregnancies, but there are differences in the cohorts considered across features set due to differences in data availability; n_{total} is the number of women (training and held-out) for each feature set. (B) Each of the three models (x-axis) and their ROC-AUC and PR-AUC (y-axes) are shown. Each of the additional EHR features performed worse than the billing codes only model and did not substantially improve performance when combined with the billing codes. Dotted lines represent chance of 0.5 for ROC-AUC and the preterm birth prevalence for PR-AUC. Even when including EHR features before and after delivery in this framework revealed the same pattern with no substantial improvement in predictive performance compared to the billing codes only model (Figure 4.9).

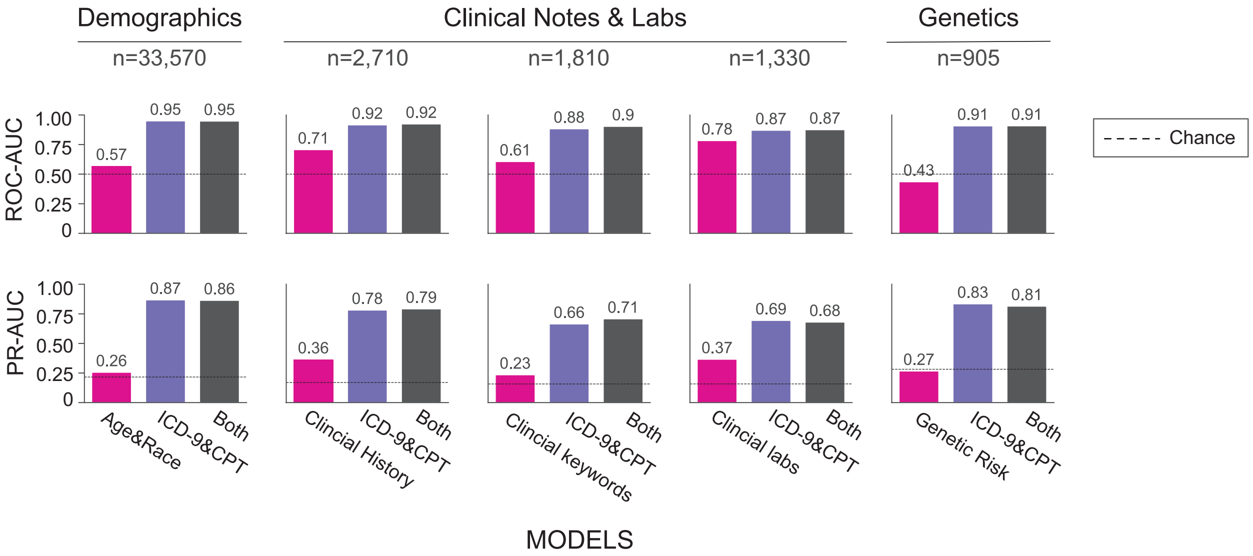


Figure 4.9**:** Combining EHR features with billing codes does not improve model performance. We evaluated how combining EHR features with billing codes could improve model performance. We used a similar framework as stated in Figure 3 but included billing codes and features before and after delivery instead of before 28 weeks of gestation. We also included multiple gestations instead of only including singleton pregnancies. EHR features are grouped in to sets of: demographic factors (age and race), clinical history (patient and familial comorbidities), clinical keywords (UMLS concept unique identifiers from obstetric notes), clinical labs, and genetic risk (polygenic risk score for preterm birth). We compared three models for each feature set: 1) using only the feature set being evaluated (pink), 2) using ICD-9 & CPT codes (purple), and 3) using the feature set combined with ICD-9&CPT codes (gray). For each feature set, we considered the subset of women who had at least one recorded value for the EHR feature and ICD&CPT codes. All three models for a given EHR feature set considered the same pregnancies, but there are differences in the cohorts considered across features set due to differences in data availability. Each of the three models (x-axis) and their ROC-AUC and PR-AUC (y-axes) are shown. Each of the additional EHR features performed worse than the billing code only model and did not substantially improve performance when combined with the billing codes. Of the other EHR features tested, clinical labs had the best predictive performance with PR-AUC of 0.37 and ROC-AUC of 0.78. Dotted lines represent chance of 0.5 for ROC-AUC and the preterm birth prevalence for PR-AUC. The total number of women (n) in each subset including the training and held-out set is given.

4.2.4 Models using billing codes outperforms prediction from risk factors

Although there are well known risk factors for preterm birth, few validated risk calculators exist and even fewer are routinely implemented in clinical practice[71]. We evaluated how a prediction model incorporating only common risk factors associated with moderate to high risk for preterm birth compared to a model using billing codes, which captured a broad range of comorbidities, at 28 weeks of gestation. We included maternal and fetal risk factors that occurred before and during the pregnancy and across many organ systems[30,31,185,199], race[182], age at delivery[200–202], pre-gestational and gestational diabetes[203], sickle cell disease[204], fetal abnormalities[30], pre-

pregnancy hypertension, gestational hypertension (including pre-eclampsia or eclampsia)[4,205], and cervical abnormalities[73] (Methods).

The billing-code-based model significantly outperformed a model trained with clinical risk factors at predicting preterm birth at 28 weeks of gestation (PR-AUC=0.40 vs. 0.25, ROC-AUC=0.75 vs. 0.65; Figure 4.10B, C). The stronger performance of the billing-code-based classifier was true for women across the spectrum of comorbidity burden; it had higher precision across individuals with different numbers of risk factors. Performance peaked for individuals with 0 (precision=0.39) and 4+ (precision=0.46) risk factors, but we did not observe a trend between model performance and increasing number of clinical risk factors (Figure 4.10D). This suggests that machine learning approaches incorporating a comprehensive clinical phenome can add value to predicting preterm birth.

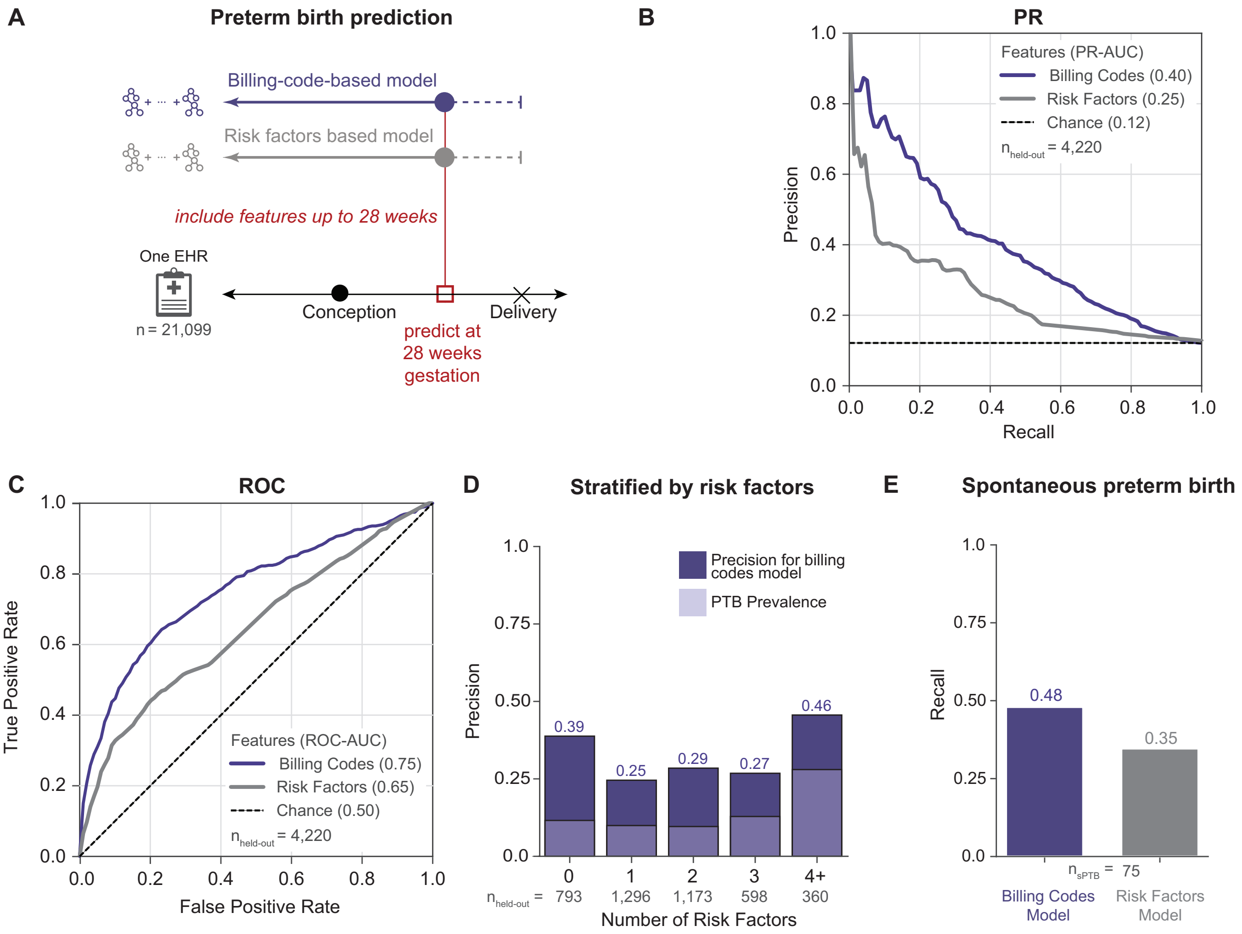


Figure 4.10: Billing-code-based model outperforms a model based on clinical risk factors. (A) We compared the performance of boosted decision trees trained using either billing codes (ICD-9 and CPT) present before 28 weeks of gestation (purple) or known clinical risk factors (gray) to predict preterm delivery. Clinical risk factors (Methods) included self- or third-party reported race (Black, Asian, or Hispanic), age at delivery (> 34 or <18 years old), non-gestational diabetes, gestational diabetes, sickle cell disease, presence of fetal abnormalities, pre-pregnancy BMI >35, pre-pregnancy hypertension (>120/80), gestational hypertension, preeclampsia, eclampsia, and cervical abnormalities. Both models were trained and evaluated on the same cohort of women (n = 21,099). (B) Precision-recall and (C) ROC curves for model using billing codes (purple line) or clinical risk factors (gray line). Preterm births are predicted more accurately by models using billing codes at 28 weeks of gestation (PR-AUC = 0.40, ROC-AUC = 0.75) than using clinical risk factors as features (PR-AUC = 0.25, ROC-AUC = 0.65). For the precision-recall curves chance performance is determined by the preterm birth prevalence (dashed black line). (D) Billing-code-based prediction model performance stratified by number of risk factors for an individual. The billing-code-based model detects more preterm cases and has higher precision (dark purple) across all numbers of risk factors compared to preterm (PTB) prevalence (light purple). (E) The model using billing codes also performs well at predicting the subset of spontaneous preterm births in the held-out set (recall = 0.48) compared to risk factors (recall = 0.35).

4.2.5 Machine learning models can predict spontaneous preterm births

The multifactorial etiologies of preterm birth led to clinical presentations with different comorbidities and trajectories. Medically-indicated and idiopathic spontaneous preterm births are distinct in etiologies and outcomes. Identifying pregnancies that ultimately result in spontaneous preterm deliveries is particularly valuable, and we anticipated that spontaneous preterm birth would be more challenging to predict than preterm birth overall. To test this, we identified spontaneous preterm births in the held-out set at 28 weeks of gestation by excluding women with medically induced labor, a cesarean section delivery, or PPROM (Methods). We intentionally used a conservative phenotyping strategy that aimed to minimize false positive spontaneous preterm births to evaluate the model's ability to predict spontaneous preterm births. The prediction model trained using billing codes up to 28 weeks of gestation classified 48% (recall) of all spontaneous preterm births (n=75) as preterm; this is significantly higher than the risk factor only model (recall = 35%; Figure 4.10E).

4.2.6 Deliveries stratifies into clusters with different risk and comorbidity signatures

Understanding the statistical patterns identified by machine learning models is crucial for their adoption into clinical practice. Unlike deep learning approaches, decision tree-based models are easier to interpret. We calculated feature importance as measured by SHapley Additive exPlanation (SHAP) scores[206,207] for each delivery and feature pair in the held-out cohort for the model using billing codes at 28 weeks of gestation ('Billing-code-based model', Figure 4.10A). SHAP scores quantify the marginal additive contribution of each feature to the model predictions for each individual. Next, we performed a density-based clustering on the patient by feature importance matrix and visualized clusters using UMAP (Figure 4.11). This approach focuses the clustering on the features for each individual prioritized by the algorithm. We identified six clusters with 927 to 102 women. Preterm birth prevalence was the high in the clusters one to four (blue, pink, green, orange) indicating differential risk for preterm birth. Performance varied across the clusters; the yellow cluster with low preterm birth prevalence had the highest PPV while clusters with higher preterm birth prevalence had higher recall.

To evaluate whether clusters had distinct phenotype profiles, we calculated the enrichment of demographic and clinical risk factor traits within each cluster using Fisher's exact test (Methods). These traits were extracted from structured fields in EHRs or ascertained using combinations of billing codes. Although these billing codes are used to train the model, the combination of codes used to ascertain risk factor traits are not encoded in the training data. White women are significantly enriched in the cluster 5 (odds ratio, OR = 1.2, p-value = 0.02, Fisher's exact test, Figure 4.11E). Hispanic women also had significant positive enrichment in cluster four (OR = 2.5, p-value = 0.0002) and cluster 6 (OR = 1.6, p-value = 0.008) and were depleted (negative enrichment) in the cluster five (OR= 0.5, p-value = 4.42E-6, Figure 4.11D). African American and Asian women also exhibit modest enrichment in different clusters.

We also tested for enrichment of clinical risk factors of preterm birth in the clusters. We observed distinct patterns of enrichment and depletion for each clinical risk factor (Figure 4.11F, Figure 4.12). Gestational hypertension had strong and enrichment in cluster three (OR = 26.4, p-value = 9.0E-39). Fetal abnormalities demonstrated a similar pattern with strong enrichment in cluster one (OR = 8.5, p-value = 2.07E-10). Extreme age at delivery (>34 or <18 years old) was enriched, though weaker, (OR = 1.2 to 2.2) for all clusters except five and six. Pre-pregnancy BMI, pre-pregnancy hypertension, and gestational hypertension had similar patterns with the strongest enrichment in cluster three. The remaining clinical risk factors show similar patterns and are provided in Figure 4.12.

By analyzing the feature importance values through UMAP embeddings, we identify interpretable clusters of individuals discovered by the machine learning model that reflect the complex and multi-faceted paths to preterm birth. Overall, the learned rules highlight relationships between clinical factors and preterm birth prevalence. For example, some risk factors, such as age at delivery, are enriched in all clusters with high preterm birth prevalence. Other factors, such as pre-pregnancy BMI and hypertension, are strongly enriched only in specific clusters with high preterm birth prevalence. Thus, this approach enables us to interpret phenotypic patterns of risk and identify subgroups among cases learned from complex EHR features by the prediction model.

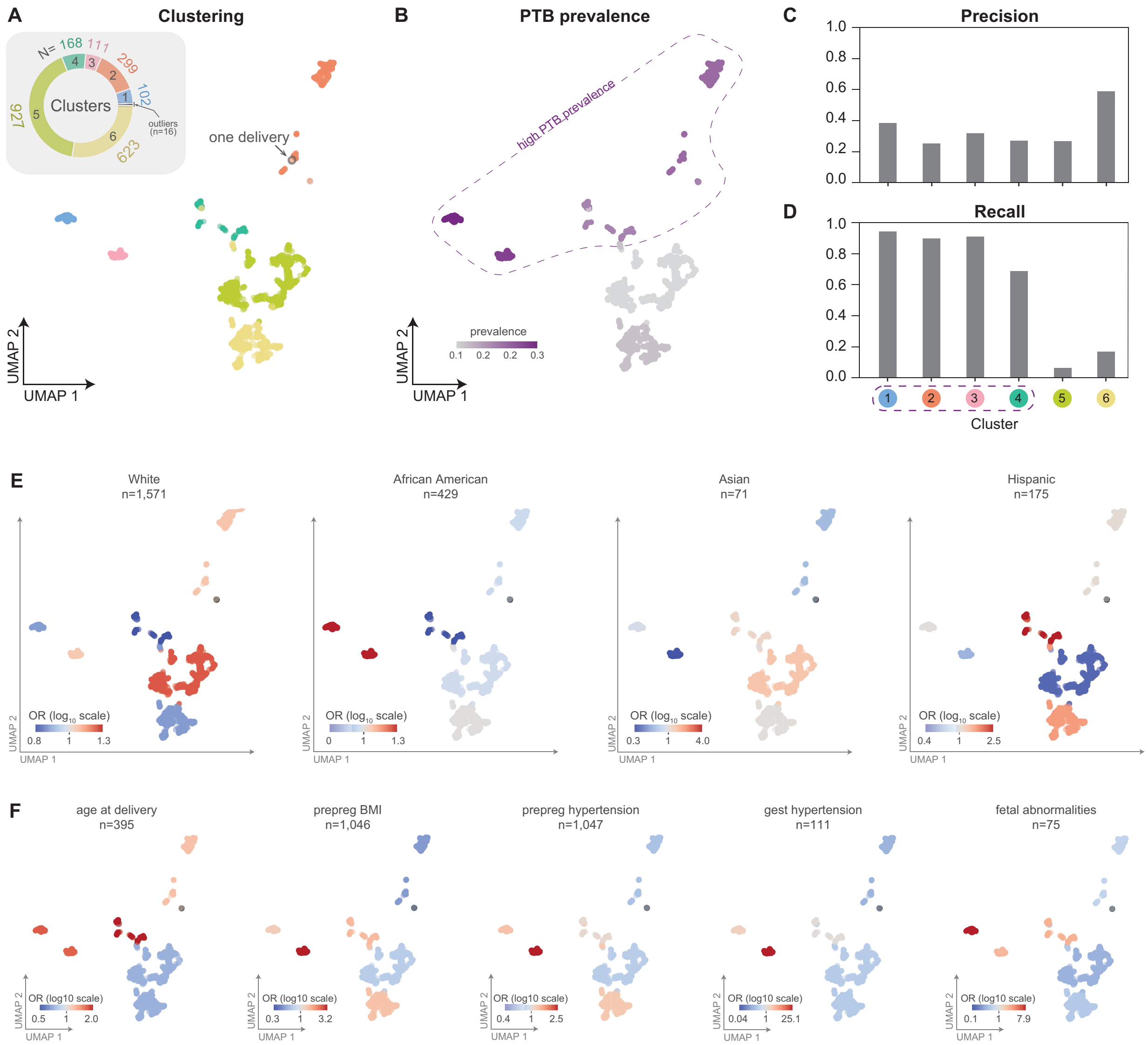


Figure 4.11: Machine-learning-based clustering of deliveries identifies sub-groups with distinct preterm birth prevalence, clinical features, and prediction accuracy. (A) For the model predicting preterm birth at 28 weeks of gestation using billing codes (ICD-9 and CPT, Figure 4.10A), we assigned deliveries from the held-out test set (n=2,246) to one of six clusters (colors) using density-based clustering (HDBSCAN) on the SHAP feature importance matrix. For visualization of the clusters, we used UMAP to embed the deliveries into a low dimensional space based on the matrix of feature importance values. Inset pie chart displays count of individuals in each cluster. (B) The preterm birth prevalence (colorbar) in each cluster. The algorithm discovered four clusters with high preterm birth prevalence (enclosed by dashed line). (C) Precision and (D) recall for preterm birth classification within each cluster. (E) The enrichment (odds ratios, colorbar in log10 scale) of race as derived from EHRs for each cluster (Table S1). (F) The enrichment ($\log_{10}$ odds ratio) of relevant clinical risk factors in each cluster. Risk factors include: age at delivery (> 34 or <18 years old), pre-pregnancy BMI (prepreg BMI), pre-pregnancy hypertension (prepreg hypertension), gestational hypertension (gest hypertension), and fetal abnormalities. We report the total number of women in the delivery cohort at high risk for each clinical risk factor (n). Enrichments for additional risk factors are given in Figure 4.12.

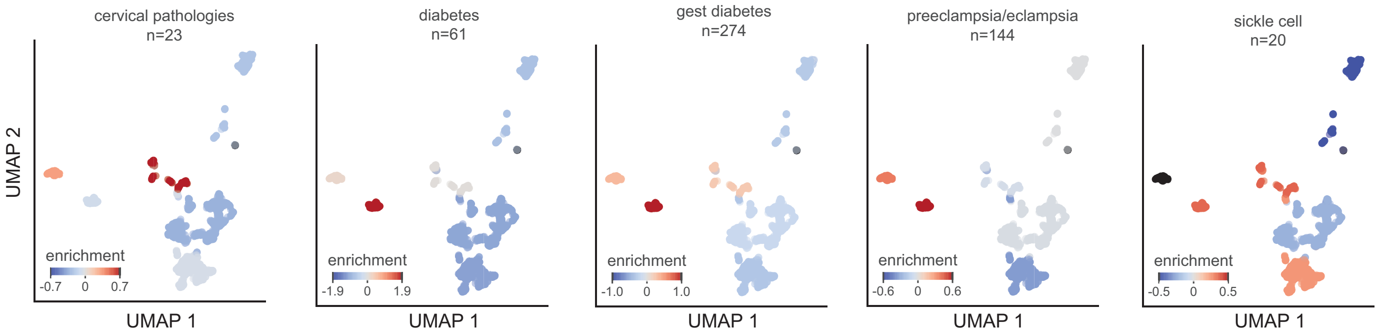


Figure 4.12: Enrichment of additional clinical risk factors in pregnancy cohort clusters. We calculated enrichment (log$_{10}$ odds ratio) of several additional clinical risk factors (each panel) for each cluster derived from the feature importance matrix for the model predicting preterm birth at 28 weeks of gestation (Figure 4.11, Methods). These risk factors are enriched in different clusters. We report the total number of women in the delivery cohort at high risk for each clinical risk factor (n).

4.2.7 Performance varies based on clinical context and delivery history

To further explore the sensitivity of the performance of our approach to clinical context and patient history, we evaluated how delivery type (vaginal vs. cesarean-section) and a previous preterm birth influence preterm birth prediction. We trained two classifiers using billing codes (ICD-9 and CPT) occurring before 28 weeks of gestation: one on a cohort of cesarean-section (n = 5,475) singleton deliveries and one on vaginal deliveries (n = 15,487). Preterm birth prediction accuracy was higher in the cesarean-section cohort (PR-AUC = 0.47, chance = 0.20) compared to the vaginal delivery cohort (PR-AUC = 0.23, chance = 0.10; Figure 4.13A). Cesarean-sections also had higher ROC-AUC compared to vaginal deliveries (0.75 vs. 0.68, Figure 4.14). As expected, the preterm birth prevalence was higher in the cesarean-section cohort.

Women with a history of preterm birth are at significantly higher risk for a subsequent preterm birth than women without a previous history. Therefore, it is particularly important to understand the drivers of risk in this cohort. We tested if models trained on EHR data of women with a history of preterm birth could accurately predict the status of their next birth. We assembled 1,416 women with a preterm birth and a subsequent delivery in the cohort and split them into a training set (80%) and held-out test set (20%) to evaluate the model performance (Methods). For these women, 53% of the second deliveries were preterm. Due to limited availability of estimated gestational age data for the recurrent preterm births, which is necessary to approximate the date of conception, we trained models using billing codes (ICD-9 and CPT) present before each of the following timepoints: 10, 30, and 60 days before the delivery. These models were all able to discriminate term from preterm deliveries better than chance (Figure 4.13B; PR-AUCs≥0.75). The model predicting a second preterm birth as early as 60 days before delivery achieved the high performance with PR-AUC=0.75 (Figure 4.13B, chance=0.53) and ROC-AUC=0.77 (Figure 4.15).

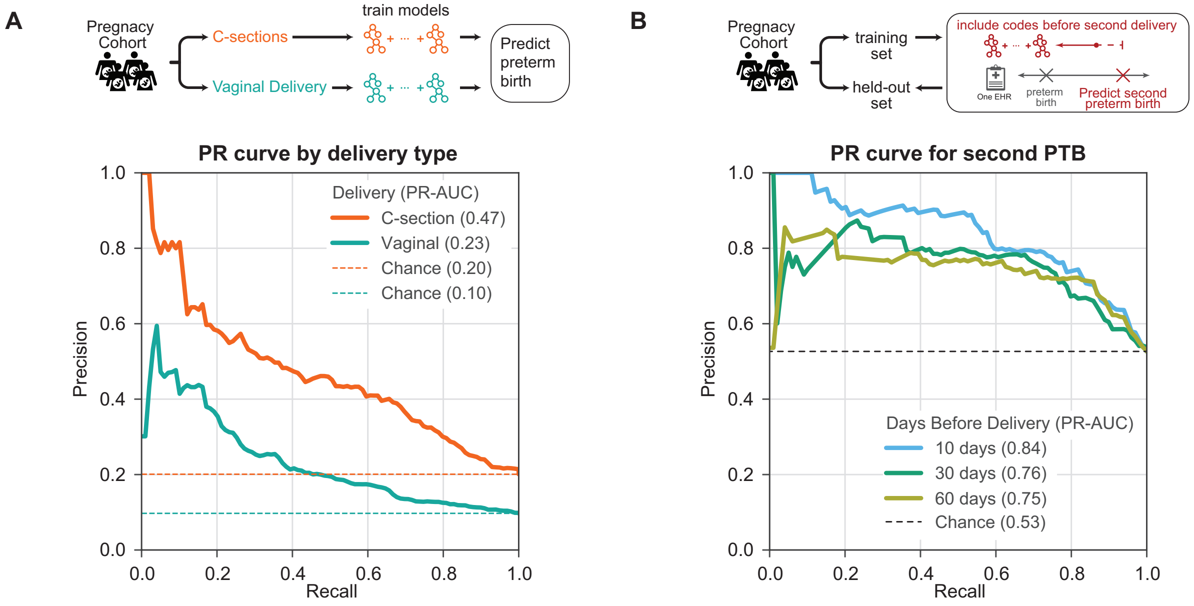


Figure 4.13: Preterm birth prediction accuracy is influenced by clinical context. (A) Preterm birth prediction models trained and evaluated only on cesarean section (C-section) deliveries perform better (PR-AUC=0.47) than those trained only on vaginal delivery (PR-AUC=0.23). ROC-AUC patterns were similar (Fig. S8). Billing codes (ICD-9 and CPT) present before 28 weeks of gestation were used to train a model to distinguish preterm from non-preterm birth for either C-sections (n=5,475) or vaginal deliveries (n=15,487). (B) Recurrent preterm birth can be accurately predicted from billing codes. We trained models to predict preterm birth for a second delivery in a cohort of 1,416 high-risk women with a prior preterm birth documented in their EHR. Three models were trained using data available at 10 days, 30 days, and 60 days before the date of second delivery. Models accurately predict the birth type in this cohort of women with a history of preterm birth (PR-AUC≥0.75). ROC-AUC varied from 0.82 at 10 days to 0.77 at 60 days before second delivery (Fig. S9). Expected performance by chance is the preterm birth prevalence in each cohort (dashed lines).

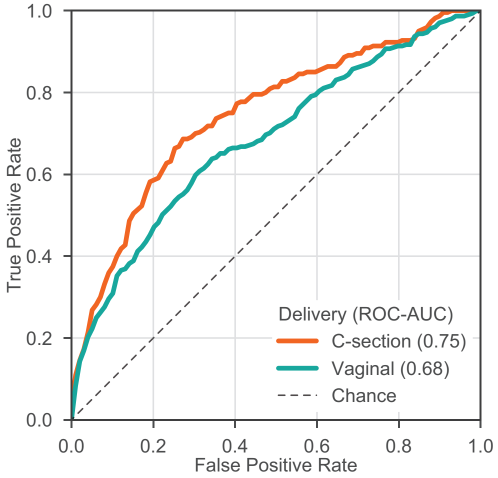


Figure 4.14: Preterm birth prediction accuracy is higher for cesarean-sections compared to vaginal deliveries. After stratifying the delivery cohort into cesarean-sections (n=5,475) and vaginal (n=15,487) deliveries, we trained a model on each delivery type to predict preterm or not-preterm births. Multiple gestations were excluded. We trained models using billing codes (ICD-9 and CPT) present before 28 weeks of gestation. ROC-AUC was higher for cesarean-sections (0.75) compared to vaginal deliveries (0.68). This corresponds to the PR curves presented in Figure 4.13.

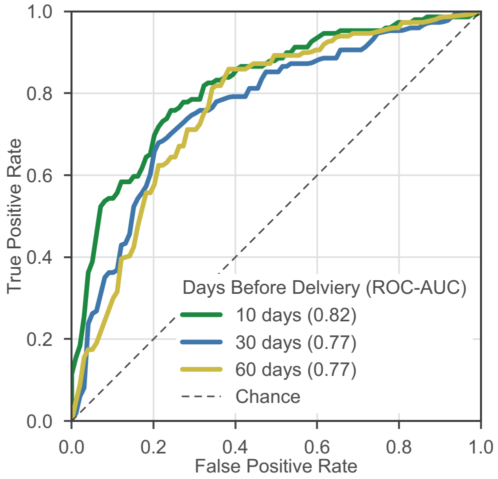


Figure 4.15: Models trained using billing codes can accurately predict risk of a second preterm birth. For women with a history of preterm birth (n=1,416, Methods), we trained models using billing codes (ICD-9 and CPT) to predict a second preterm birth. Multiple gestations were excluded. For each model, only billing codes timestamped before the specified number of days before delivery are included. Models predicted a second preterm birth accurately with the highest and lowest ROC-AUC of 0.82 at 10 days and 0.77 at 60 days before delivery respectively. This corresponds to the PR curves presented in Figure 4.13B.

4.2.8 *Models trained at Vanderbilt accurately predict preterm birth in an independent cohort at UCSF*

To evaluate whether preterm birth prediction models trained on the Vanderbilt cohort performed well on EHR data from other databases, we compared their performance on the held-out Vanderbilt cohort (n=4,215) and an independent cohort from UCSF (n=5,978). The UCSF cohort was ascertained using similar rules as the Vanderbilt cohort (Methods); age and distribution of race are provided in Table 4.1. However, we note that the UCSF cohort has a lower preterm birth prevalence (6%) compared to the Vanderbilt cohort (13%).

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

Table 4.1: Demographic distribution of UCSF and Vanderbilt cohorts. We identified women with preterm and not preterm deliveries at UCSF and Vanderiblt using similar ascertainment (Methods). For each woman, we predicted the earliest delivery in their EHR. We report age at delivery (Patient Age) and self- or third-party reported race for both cohorts. T-tests and chi-squared tests of independence were used to compare distributions stratified by delivery label.

To facilitate the comparison, we trained models to predict preterm birth in the Vanderbilt cohort using only ICD-9 codes present before 28 weeks of gestation. We did not consider CPT codes in this analysis due to differences in the available billing code data between Vanderbilt and UCSF. As expected from the previous results, the model accurately predicted preterm birth in the held-out set from Vanderbilt (PR-AUC of 0.34, chance=0.12), but performance was slightly lower than using both ICD and CPT codes (Figure 4.10). The model trained at Vanderbilt also achieved strong performance in the UCSF cohort. The classifier had a higher ROC-AUC (0.80) in UCSF cohort compared to the Vanderbilt cohort (0.72; Figure 4.16A) and PR-AUC of 0.31 vs 0.34 at Vanderbilt; Figure 4.16B). The higher ROC is due to the lower prevalence of preterm birth in the UCSF cohort and the sensitivity of ROC-AUC to class imbalance[208]. Overall, these models show striking reproducibility across two independent cohorts.

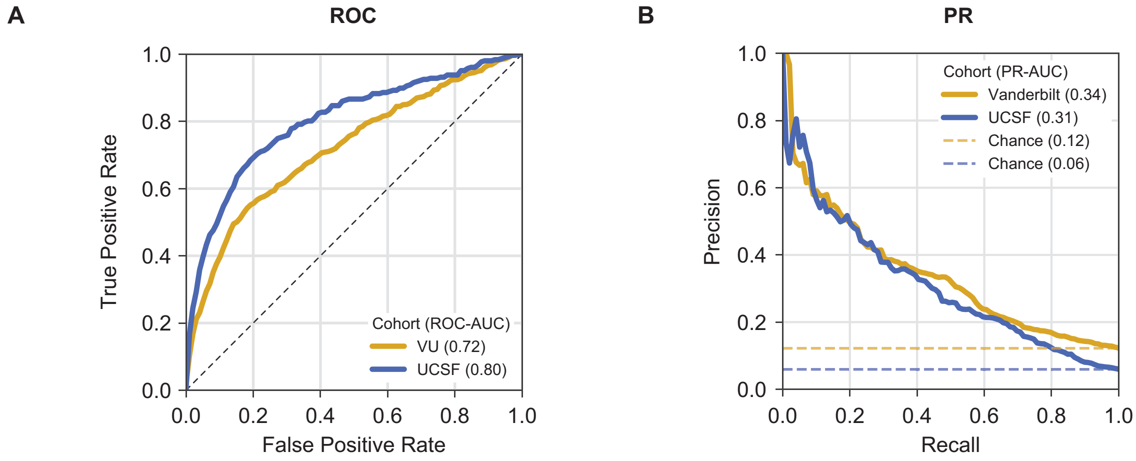


Figure 4.16: Preterm birth prediction models accurately generalize to an independent cohort. Performance of preterm birth prediction models trained at Vanderbilt applied to UCSF cohort. Models were trained using ICD-9 codes present before 28 weeks of gestation at Vanderbilt on 16,857 of women and evaluated on a held-out set at Vanderbilt (n=4,215, gold) and UCSF cohort (n=5,978, blue). (A) Models accurately predicted preterm birth at Vanderbilt (ROC-AUC=0.72) and at UCSF (ROC-AUC=0.80). The higher ROC-AUC at UCSF is driven by the lower prevalence of preterm birth in this cohort. (B) Models performed better than baseline prevalence (chance) based on the precision-recall curve at Vanderbilt (PR-AUC=0.34) and at UCSF (PR-AUC=0.31). Note that in contrast to models presented previously this one was trained only on ICD-9 codes, due to the lack of CPT codes in the UCSF cohort. Feature importance estimates were strongly correlated between the two cohorts (Figure 4.17). Cohort demographics are given in Table 4.1.

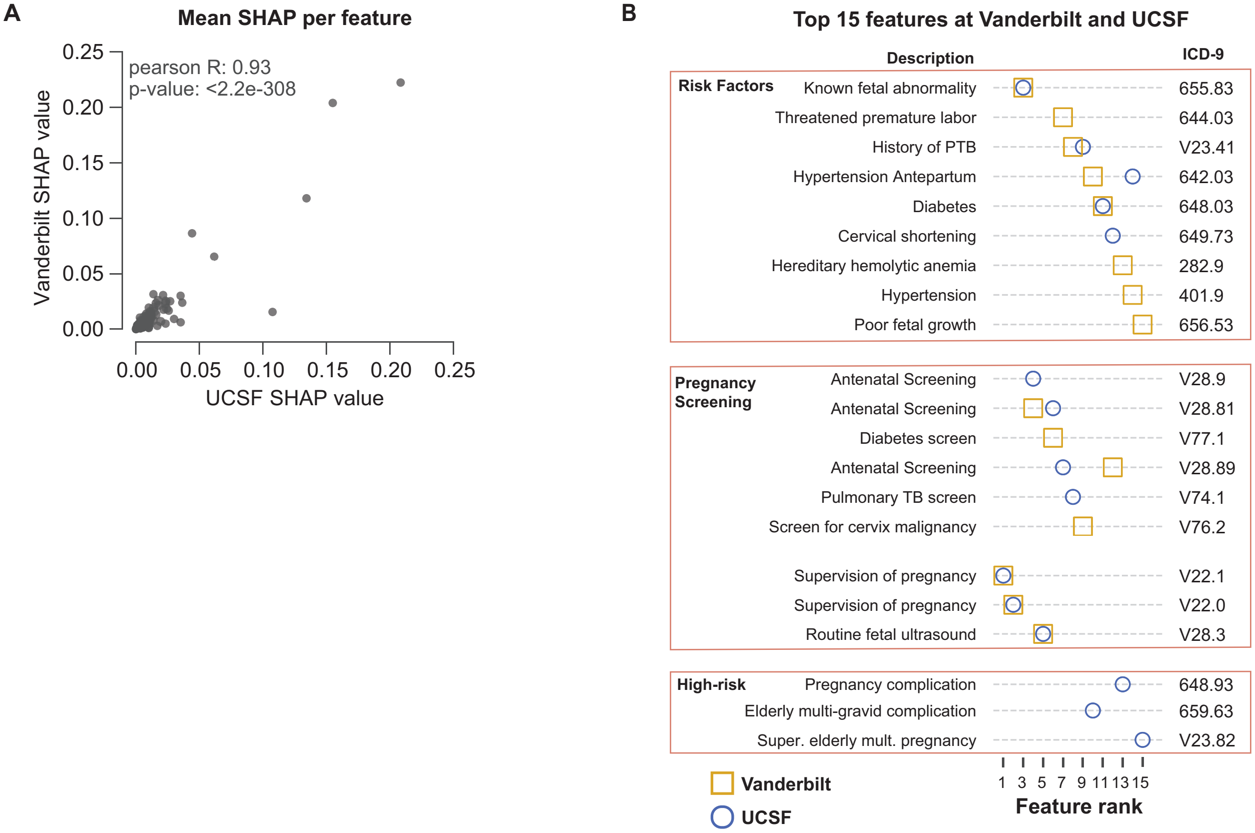


Figure 4.17: Preterm birth model feature importance is similar in an external cohort. A preterm birth prediction model trained at Vanderbilt was applied to an external UCSF cohort. Models were trained using ICD-9 codes present before 28 weeks of gestation at Vanderbilt on 16,857 of women and evaluated on a held-out set at Vanderbilt (n=4,215, gold) and UCSF cohort (n=5,978, blue). These models performed similarly (Figure 4.16). (A) Feature importance was estimated by the mean absolute SHapley Additive exPlanation (SHAP) value per feature in each individual in each cohort (x and y-axes). The feature importance estimates have a high positive correlation between cohorts (Pearson r=0.93, p<2.2e-308, two-tailed). (B) The top 15 features with the highest mean absolute SHAP score in the Vanderbilt cohort (gold square) or UCSF cohort (blue circle). The majority of the features were shared across cohorts and capture known risk factors (fetal abnormalities, history of preterm birth, etc.), pregnancy screening visits, and supervision of high-risk pregnancies.

4.2.9 Similar features are predictive across the independent cohorts

The architecture of boosted decision trees enables straightforward identification of features (ICD-9 codes) with the largest influence on the model predictions. We used SHAP[209,210] scores to quantify the marginal additive contribution of each feature to the model predictions for each individual. For each feature in the ICD-9-based model, we calculated the mean absolute SHAP values across all women in the held-out set. The mean absolute SHAP value for each feature was highly correlated (spearman R=0.93, p-value < 2.2E-308) between the held-out Vanderbilt set and the UCSF cohort (Figure 4.17A). The top 15 features ranked based on the mean absolute SHAP value captured known risk factors (fetal abnormalities, history of preterm birth, etc.), pregnancy screening and supervision of high-risk pregnancies (Figure 4.17B). Ten of the top 15 features were shared across both cohorts. This suggests that the model's discovered combination of phenotypes, including expected risk factors, and the corresponding weights assigned by the machine learning model are generalizable across cohorts.

4.3 Discussion

Preterm birth is a major health challenge affecting ~10% of pregnancies[2,4,6] and lead to significant morbidity and mortality[211,212]. Predicting preterm birth risk could inform clinical management, but no accurate classification strategies are routinely implemented[186]. Here, we take a step toward addressing this need by demonstrating the potential for machine learning on dense phenotyping from EHRs to predict preterm birth in challenging clinical contexts (e.g., spontaneous and recurrent preterm births). However, we emphasize that more work is needed before these approaches are ready for the clinic. Compared to other data types in the EHRs, models using billing codes alone had the highest prediction accuracy and outperformed those using clinical risk factors. Demonstrating the potential broad applicability of our approach, the model accuracy was similar in an external independent cohort. Combinations of many known risk factors and patterns of care drove prediction; this suggests that the algorithm builds on existing knowledge. Thus, we conclude that machine learning based on EHR data has the potential to predict preterm birth accurately across multiple healthcare systems.

Decision tree-based models are robust to correlated features, can identify complex non-linear combinations, and remain transparent for interpretation after training. In addition to these advantages, decision tree-based models have demonstrated strong performance in various clinical prediction tasks[213–215]. Pregnancy is a clinical context with close monitoring and well defined end-points that may similarly benefit from machine learning approaches, yet few studies have applied decision tree based machine learning models to large pregnancy cohorts with rich clinical data[216].

Our approach has several distinct advantages compared to published preterm birth prediction models. First, our models have robust performance. Previous models using risk factors (diabetes, hypertension, sickle cell disease, history of preterm birth) to predict preterm birth, despite having cohorts up to two million women[185], have reported ROC-AUCs between 0.69 and 0.74[182–184]. Our models obtain a ROC-AUC of 0.75 and PR-AUC of 0.40 using data available at 28 weeks of gestation even after excluding multiple gestations. Furthermore, given the unbalanced classification problem (preterm births are less common than non-preterm), we report high PR-AUCs in addition to high ROC-AUCs. A recent deep learning model trained using word embeddings from EHRs achieved a high performance (ROC-AUC = 0.83[216]). This model was evaluated over a stratified high-risk cohort consisting of birth before 28 weeks of gestation. We did not stratify preterm births by severity since more than 85% of preterm births occur after 32 weeks of gestation[217], however, this is an important topic for future work. Our models achieve comparable performance with the benefit of easier interpretability, which is an advantage over deep learning approaches, and we discuss this further below.

Second, our models use readily available data throughout pregnancy that do not require invasive sampling. While some studies have also obtained high ROC-AUCs (e.g., 0.81-0.88), they used serum biomarkers across small cohorts[181] or acute obstetric changes within days of delivery[180]. The potential to enable cost-effective and broad application is illustrated by our evaluation of the classifiers on EHR data from UCSF; however, substantial further work is needed to move from this proof-of-concept analysis to clinic-ready models. Furthermore, the rich characterization of the phenome provided by EHRs leveraged by our approach could also complement more invasive biochemical assays.

Third, the gradient boosted decision trees we implement are easier to interpret than ‘black-box’ deep learning models that cannot easily identify features driving predictions. Transparency is an important, if not necessary, characteristic of machine/artificial learning models deployed in clinical practice[218,219], and it can facilitate discovery of insights and hypotheses to motivate future work. We reveal the patterns learned by our model by clustering deliveries using feature importance profiles. The enrichment for known risk factors in clusters with high preterm birth prevalence establishes confidence in our machine-learning based prediction models. In addition, we can quantify the strength of enrichment and combination of risk factors across clusters with distinct comorbid patterns. Since preterm birth is a heterogenous phenotype[49], and stratifying pregnancies based on clinical features may be critical to uncovering the biological basis of labor[31,58,61], the learned rules from our model offer a possible method for sub-phenotyping.

Finally, our approach generalizes across hospital systems. We demonstrate that billing-code-based models trained at Vanderbilt achieve similar accuracy in an independent cohort from UCSF. The generalizability of machine learning models can be constrained by the sampling of the training data. Thus, the accurate prediction in an independent dataset from an external institution points to several inherent strengths of the approach. First, successful replication indicates the models' ability to learn predictive signals despite regional variation in assigning billing codes to an EHR. Second, the large cohorts used to train and evaluate models at Vanderbilt and USCF guard against potential weakness of EHRs, such as miscoding or omission of key data points. These errors are unavoidable in EHRs[220], but the large cohort used to train our models mitigates these errors and enables the high accuracy in the UCSF dataset, even with its different demographics. Additionally, idiosyncratic patterns of patient care at the institution used to develop the algorithm, which would be present in the Vanderbilt training and held-out sets, are unlikely to be present in the external UCSF cohort and inflate the out-of-sample accuracy. Third, the top features driving model performance are shared across institutions and reflect combinations of known risk factors and patterns of care. This aids interpretability of the underlying algorithm and likely reflects underlying pathophysiology that is innate to preterm birth.

We see several avenues for further improving our algorithm. First, some of the top features reflected routine obstetric care for high-risk pregnancies. Thus, the learning problem could be engineered to force the algorithm to discover new unappreciated risk factors. Second, we were surprised that the addition of features beyond billing codes, such as lab values, concepts extracted from clinical notes, and genetic information did not significantly improve performance. In some cases, any redundant information already captured by the billing codes would not improve the model's accuracy; this is likely true for clinical notes. However, other sources, like currently available genetic data and polygenic risk scores, may not effectively capture underlying etiologies of preterm birth. Thus, these sources may not add more discriminatory power due to limitations in current data. Indeed, the largest published genome-wide study for preterm birth only explains a very small fraction of the heritability[47], and a polygenic risk score derived from it was not predictive in our cohort. Further sub-phenotyping of preterm birth will not only aid in prediction, but also understanding its multifactorial etiology and developing personalized treatment strategies. More explicit modeling of the temporal dependence between EHR features may further increase performance. Finally, while we evaluated the ability of our classifiers to discriminate preterm births, further studies evaluating the calibration of these models are necessary to better risk stratify of pregnancies.

The strong predictive performance of our models suggests that they have the potential to be clinically useful. Compared to a machine learning model trained using only known risk factors, the billing-code-based classifier incorporated a broad set of clinical features and predicted preterm birth with higher accuracy. Furthermore, the superior performance was not driven by the number of risk factors or the total burden of billing codes. These results indicate the algorithm is not simply identifying less healthy individuals or those with greater healthcare usage. The models also accurately predicted many preterm births in challenging and important clinical contexts such as spontaneous and recurrent preterm birth. Spontaneous preterm births are common[4,6,56], and unlike iatrogenic deliveries, they are more difficult to predict because they are driven by unknown multifactorial etiologies[6,186]. Similarly, since a prior history of preterm birth is one of the strongest risk factors[221], distinguishing pregnancies most at risk for recurrent preterm birth has potential to provide clinical value.

However, we emphasize that additional work is needed before this approach is ready for clinical application. Though it has strong performance, a more comprehensive evaluation of the algorithm against current clinical practice is needed to determine how early and how much improvement in standard of care this approach could provide[222]. Furthermore, while our cohorts include diverse individuals and the algorithm generalizes well, the approach must be evaluated to ensure that it does not introduce of amplify biases against specific groups or types of preterm birth[223]. In addition, we anticipate further gains in the clinical value of this approach as more modalities of data becomes incorporated in the EHR[224] and diverse populations become available. Addressing these questions and taking other necessary steps toward clinical utility will require the close collaboration of diverse experts from basic, clinical, social, and implementation sciences.

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

4.4 Methods

4.4.1 Ascertaining delivery type and date for Vanderbilt cohort

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

We demarcate multiple deliveries by grouping delivery-ICDs in intervals of 37 weeks starting with the most recent delivery-ICD. This step is repeated until all delivery-ICDs in a patient's EHR are assigned to a pregnancy. We chose 37-week intervals to maximally discriminate between pregnancies. For each delivery, we assign a list of labels (preterm, term, or postterm) ascertained using the delivery-ICDs. EGA values, extracted from structured fields across clinical notes, were mapped to multiple pregnancies using the same procedure. For women with multiple EGA recorded in their EHR, the most recent EGA value determined the time interval to group preceding EGA values. Based on the most recent EGA value for each pregnancy, we assigned labels to each delivery (EGA <37 weeks: preterm; ≥37 and <42 weeks: term, ≥42 weeks: postterm). After pooling delivery labels based on delivery-ICDs and EGA, we assigned a consensus delivery label by selecting the oldest gestational age-based classification (i.e. postterm > term > preterm). By incorporating both billing code and EGA based delivery label and selecting the oldest gestational classification, we expect this to increase the accuracy of this algorithm, which we evaluate by chart-review (described in detail below).

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

4.4.2 Validating delivery type based on chart review

To validate the delivery type ascertained from billing codes and EGA, we used chart-reviewed labels as the gold standard. For 104 randomly selected EHRs from the delivery cohort, we extracted the date and gestational age at delivery from clinical notes. For earliest delivery recorded in the EHR, we assigned a chart-review based label according to the gestational age at delivery (<37 weeks: preterm; 37 and 42 weeks: term, ≥42 weeks: post term). The precision/positive predictive value for the ascertained delivery type as a binary variable (‘preterm’ or ‘not-preterm’) was calculated using the chart reviewed label as the gold standard. To compare the ascertainment strategy to a simpler phenotyping algorithm, we compared the concordance of the label derived from delivery-ICDs to one based on the gestational age within three days of delivery. This simpler phenotyping approach resulted in a lower PPV (85%) and recall (93%; Figure 4.1) compared to the billing-code-based ascertainment strategy.

4.4.3 Training and evaluating gradient boosted decision trees to predict preterm birth

All models for predicting preterm birth used boosted decision trees as implemented in XGBoost v0.82[195]. Unless stated otherwise, we trained models to predict the earliest delivery in a woman’s EHR as preterm or not-preterm. The delivery cohort was randomly split into training (80%) and held-out (20%) sets with equal proportion of preterm cases. For prediction tasks, we used only ICD-9 and excluded ICD-10 codes to avoid potential confounding effects. The total count of billing codes within a specified time frame was used as features to train our models; if a woman never had a billing code in her EHR, we encoded these as ‘0’. For all models we excluded ICD-9, CPT codes, and EGA used to ascertain delivery type and date. On the training set, we use tree of Parzen estimators as implemented in hyperopt v0.1.1[225] to optimize hyperparameters by maximizing the mean average precision. The best set of hyperparameters was selected after 1,000 trials using 3-fold cross-validation over the training set (80:20 split with equal proportion of preterm cases). We evaluated the performance of all models on the held-out set using Scikit-learn v0.20.2[226]. All performance metrics reported are on the held-out set. For precision-recall curves, we define baseline chance for each model as the prevalence of preterm cases. To ensure no data leaks were present in our training protocol, we trained and evaluated a model using a randomly generated dataset (n=1,000 samples) with a 22% preterm prevalence. As expected, this model did not do better than chance (AUC=0.50, PR-AUC=0.22, data not show). All trained models with their optimized hyperparameters are provided at https://github.com/abraham-abin13/ptb_predict_ml.

4.4.4 Predicting preterm birth at different weeks of gestation

As a first step, we evaluated whether billing codes could discriminate between delivery types. Models were trained to predict preterm birth using the total counts of each ICD-9, CPT, or ICD-9 and CPT code across a woman's EHR. We excluded any codes used to ascertain delivery type or date. All three models were trained and evaluated on the same cohort of women who had at least one ICD-9 and CPT code (Figure 4.4).

Next, we evaluated machine learning models at 0, 13, 28, and 35 weeks of gestation by training using only features present before each timepoint. For the subset of women in our delivery cohort with EGA, we calculated the date of conception by subtracting EGA (recorded within three days of delivery) from the date of delivery. Next, we trained models using ICD-9 and CPT codes timestamped before different gestational timepoints with only singleton (Figure 4.3) or including multiple gestations (Figure 4.5). The same cohort of women was used to train and evaluate across models. The sample size varied slightly (n = 11,843 to 10,799) since women who already delivered were excluded at each timepoint.

In addition to evaluating models based on the date of conception, we trained models at different timepoints before the date of delivery (Figure 4.6) using the same cohort of women by requiring every individual in this cohort had to have at least one ICD-9 or CPT code before each timepoint. Evaluating models before the date of delivery increased the sample size (n=15,481) compared to a prospective conception-based design (n=12,410) and yielded similar results.

4.4.5 Evaluating predictive potential of demographic, clinical, and genetic features from EHRs

In addition to billing codes, we extracted structured and unstructured features from the EHRs (Figure 4.8A). We evaluated models using features present before 28 weeks of gestations (Figure 4.8) and features present before or after delivery (Figure 4.9). Structured data included self or third-party reported race (Figure 4.1E), age at delivery, past medical and family history (92 features), and clinical labs. For training models, we only included clinical labs obtained during the first pregnancy and excluded values greater than four standard deviations from the mean. To capture the trajectory of each clinical lab's values across pregnancy we trained models using the mean, median, minimum, and maximum lab measurement. For unstructured clinical text in obstetric and nursing clinical notes, we applied CLAMP[227] to extract UMLS (Unified Medical Language System) concepts unique identifiers (CUIs and included those with positive assertions with > 0.5% frequency across all EHRs). When training preterm birth prediction models, we one-hot encoded categorical features. No transformations were applied to the continuous features.

A subset of women (n=905) was genotyped on the Illumina MEGA^{EX} platform. We applied standard GWAS quality control steps[228] using PLINK v1.90b4s[85]. We calculated a polygenic risk score for each white woman with genotype data based on the largest available preterm birth GWAS [47] using PRSice-2[229,230]. We assumed an additive model and summed the number of risk alleles at single nucleotide polymorphisms (SNPs) weighted by their strength of association with preterm birth (effect size). PRSice determined the optimum number of SNPs by testing the polygenic risk score for association with preterm birth in our delivery-cohort at different GWAS p-value thresholds. We included date of birth and five genetic principal components to control for ancestry. Our final polygenic risk score used 356 preterm birth associated SNPs (GWAS p-value < 0.00025).

Using the structured and unstructured data derived from the EHR, we evaluated whether adding EHR features to billing codes could improve preterm birth prediction. Since the number of women varied across EHR feature, we created subsets of the delivery cohort for each EHR feature. Each subset included women with at least one recorded value for the EHR feature and billing codes. Then we trained three models as described above for each subset: 1) using only the EHR feature being evaluated, 2) using ICD-9 & CPT codes, and 3) using the EHR feature with

ICD-9 & CPT codes. Thus, all three models for a given EHR feature were trained and evaluated on the same cohort of deliveries (Figure 4.8).

4.4.6 Predicting preterm birth using billing codes and clinical risk factors at 28 weeks of gestation

We compared the performance of a model trained using billing codes (ICD-9 and CPT) present before 28 weeks of gestation with a model trained using clinical risk factors to predict preterm delivery (Figure 4.10). Both models were trained and evaluated on the same cohort of women (n = 21,099). We selected well-established obstetric risk factors that included maternal and fetal factors across organ systems, occurred before and during pregnancy, and had moderate to high risk for preterm birth [30,31,185,199]. For each individual, risk factors were encoded as high-risk or low-risk binary values. Risk factors such as non-gestational diabetes status[203], gestational diabetes[203], gestational hypertension, pre-eclampsia or eclampsia[4,205], fetal abnormalities[30], cervical abnormalities[73], and sickle cell disease[204] status was defined based on at least one corresponding ICD-9 code occurring before the date of delivery. The remaining factors, such as race (Black, Asian, or Hispanic was encoded as higher risk)[182], age at delivery (> 34 or <18 years old)[200–202], pre-pregnancy BMI ≥ 35, and pre-pregnancy hypertension (>120/80)[4,205], were extracted from structured fields in EHR. Pre-pregnancy value was defined as the most recent measurement occurring before nine months of the delivery date.

4.4.7 Density based clustering on feature importance values

To better understand the decision making process of our machine learning models, we calculated feature importance value for the model predicting preterm birth at 28 weeks of gestation. We used SHapley Additive exPlanation values (SHAP)[206,207,210] to determine the marginal additive contribution of each feature for each individual. First, we calculated a matrix of SHAP values of features by individuals from the held-out cohort. Since the shape of this matrix was too large to perform the density based clustering, we created an embedding using 30 UMAP components with default parameters as implemented in UMAPv0.3.8[231]. Next, we performed a density based hierarchical clustering using HDBSCANv0.8.26 [232]. We used default parameters (metric=Euclidean) and tried a range of values for two hyperparameters: minimum number of individuals in each cluster (‘min_clust_size) and threshold for determining outlier individuals who do not belong to a cluster (‘min_samples’). After tuning these two hyperparameters, we selected the clustering model with the highest density based cluster validity score [232], which measures the within and between cluster density connectedness. We find a min_clust_size = 110 and min_samples = 10 had the highest density based cluster validity (DBCV) score with 6 distinct clusters with one cluster for outliers (Figure 4.18). A minority of women (n=16) were not assigned to a cluster (‘outliers’). To visualize the cluster assignments, we performed UMAP on the feature importance matrix with default settings and two UMAP components and colored each individual by their cluster membership. Finally, we calculated the preterm birth prevalence and accuracy within each cluster.

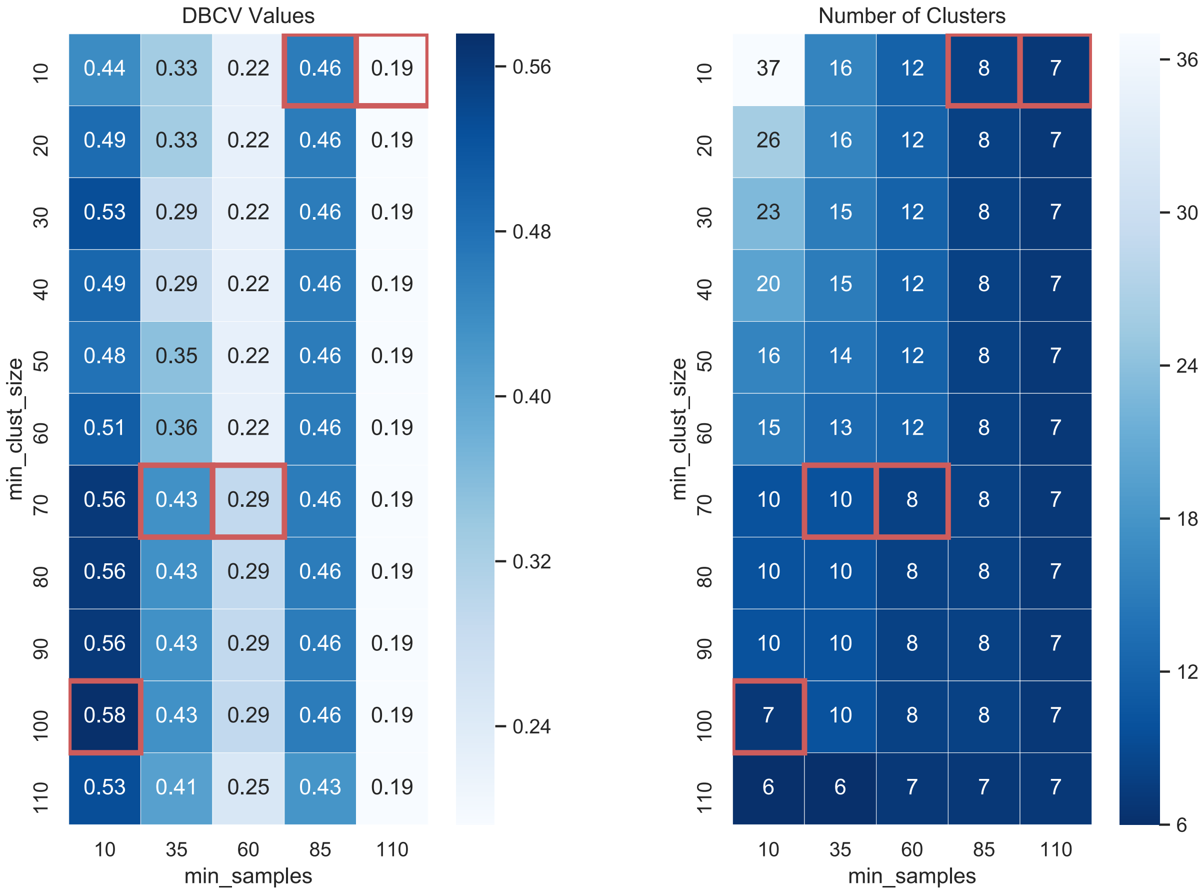


Figure 4.18: Density based cluster validity score across hyper-parameters space for HDBSCAN clustering of deliveries by feature importance. To identify the optimum number of clusters using HDBSCAN on the billing-code-based model at 28 weeks gestation in the held-out set, we explored two hyperparameters: minimum number of individuals in each cluster (‘min_clust_size, y-axis) and threshold for determining outlier individuals who do not belong to a cluster (‘min_samples’, x-axis). The left heatmap represents cluster validity measured with the density-based cluster validity (DBCV) score with higher DBCV (darker blue) scores indicating more distinct clusters. The right heatmap displays the number of clusters (ligher blue == higher number of clusters) for the pair of hyperparameters. Cells outlined in red have the highest values within their column. Note, number of clusters includes a cluster for outliers.

4.4.8 Comorbidity enrichment within clusters

We tested for enrichment of clinical risk factors within each cluster by using a Fisher Exact test as implemented in Scipy[233]. For each risk factor, we constructed a contingency table based on a given cluster membership and being high risk for the risk factor. We report enrichment as the odds ratio with colorbar in log10 scale of the odds ratio. For sickle cell disease, one cluster did not have any cases of sickle cell disease.

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

We compared how models trained used billing codes (ICD-9 & CPT) performed in different clinical contexts. First, we evaluated the accuracy of predicting spontaneous preterm birth using models trained to predict all types of preterm births. From all preterm cases in the held-out set, we excluded women who met any of the following criteria to create a cohort of spontaneous preterm births: medically induced labor, delivery by cesarean section, or preterm premature rupture of membranes. The ICD-9 and CPT codes used to identify exclusion criteria are provided in Supplementary Materials and Methods. We calculated recall/sensitivity as the number of predicted spontaneous preterm births out of all spontaneous preterm births in the held-out set. We used the same approach to quantify performance of models trained using clinical risk factors (Figure 4.10).

We trained models to predict preterm birth among cesarean sections and vaginal deliveries separately using billing codes (ICD-9 & CPT) as features. Deliveries were labeled as cesarean sections or vaginal deliveries if they had at least one relevant billing code (ICD-9 or CPT) occurring within ten days of the date of first delivery in EHR. Billing codes used to determine delivery type are provided in Supplementary Materials and Methods. Deliveries with billing codes for both cesarean and vaginal deliveries were excluded. We trained separate models to predict cesarean and vaginal deliveries (Figure 4.13).

We evaluated how well models using billing codes could predict recurrent preterm birth. From our delivery cohort, we retained women whose first delivery in the EHR was preterm and a second delivery for which we ascertained the type (preterm vs. not-preterm) as described above for the first delivery. We trained models using billing codes (ICD-9 & CPT) at timepoints before the date of delivery because the majority of this cohort did not have reliable EGA at the second delivery. As described earlier, separate models were trained using billing codes timestamped before timepoint being evaluated (Figure 4.13, Figure 4.15).

4.4.10 Preterm birth prediction in independent UCSF cohort

We evaluated how well models trained at Vanderbilt using billing codes would replicate in an external cohort assembled at UCSF. Only the first delivery in the EHR was used for prediction.

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

4.4.11 Feature interpretation from boosted decision tree models

To determine feature importance, we used SHapley Additive exPlanation values (SHAP)[207,209,210] to determine the marginal additive contribution of each feature. For the held-out Vanderbilt cohort and the UCSF cohort, a SHAP value was calculated for each feature per individual. Feature importance was summarized by taking the mean of the absolute value of SHAP scores across individuals. The top fifteen features based on the mean absolute SHAP value in either the Vanderbilt or UCSF cohorts values are reported. To compare how feature importance varies at Vanderbilt and UCSF, we computed the Pearson correlation of the mean absolute SHAP values.

CHAPTER V

5 Conclusions and Future Directions

5.1 Summary

This dissertation advances our understanding of the phenotypic, genetic, and evolutionary heterogeneity of preterm birth. Using dense and longitudinal data extracted from EHRs, I used unsupervised methods to identify distinct sub-phenotypes of preterm birth. I then applied tensor decomposition and identified latent factors, representative of sub-phenotypes, in a large cohort of Black and White women with preterm or term delivery. Further downstream analyses on these latent factors revealed a subset of women with increased polygenic burden for comorbidities known to increase preterm birth risk. Next, I employed an evolutionary perspective to demonstrate how genomic regions associated with preterm birth have been shaped by a diverse set of evolutionary forces. The evolutionary analysis highlighted preterm birth genomic regions that did not meet genome-wide significance and are high-priority candidates for future studies. For the final aim, I combined both genetic and non-genetic features extracted from EHRs to create a robust machine learning predictor of preterm birth. This model was rigorously evaluated across clinical contexts and validated in an external cohort. This work serves as a proof-of-concept that machine learning models derived from EHRs can assist in medical management and improve maternal health. In the following sections, I describe key conclusions and future directions.

5.2 Unsupervised sub-phenotyping of heterogenous traits like preterm birth can accelerate GWAS discoveries

5.2.1 Genome-wide association studies have had limited success in preterm birth

Genome-wide association studies (GWAS) have had tremendous success expanding the map of genotype to phenotype relationships across many traits. Large genetic databases, such as the UK Biobank, have accelerated the number of genome-wide association studies by providing unprecedented access to large cohorts across a broad spectrum of phenotypes. However, this success has not been universal across all disease traits. Many diseases, including preterm birth, are heterogenous with varied clinical presentations, associated comorbidities, and multifactorial etiologies. Isolating specific mechanisms in a heterogenous cohort is challenging and so far only a few genomic regions have been robustly associated with preterm birth[47].

5.2.2 Phenotyping algorithms can efficiently scale to generate large cohorts for genome-wide association studies of preterm birth

One approach to studying heterogenous traits is to subset cases based on a specific disease feature. The inherent challenge in creating subsets of a heterogenous disease is lower statistical power as a result of having smaller sample sizes. For example, the largest genome-wide association study of tens of thousands of European women with preterm birth included only spontaneous deliveries. Nevertheless, the cohort size of this study pales in comparison to other well-studied traits where the sample size numbers in the millions. For preterm birth, linking of de-identified EHRs to genetic biobanks enables rapid ascertainment of cases and controls for genome-wide association studies. As an example of a phenotyping algorithm derived from EHRs[234,235], in chapter four, I developed an algorithm to identify deliveries using multiple billing codes and estimated gestational age. Validation by chart review demonstrated the high accuracy to ascertain preterm birth cases and controls. Although the accuracy will have to be re-evaluated in new biobanks, one advantage of this phenotyping algorithm is its portability to new health systems. Even for uncommon traits such as spontaneous preterm birth, aggregating cases across multiple large genetic biobanks, which themselves have hundreds of thousands of individuals, can quickly assemble a large cohort for genome-wide association studies.

5.2.3 Unsupervised phenotyping can identify unbiased sub-phenotypes with potentially distinct etiologies

Unlike validating phenotyping algorithms using domain specific knowledge as markers for ascertainment, unsupervised phenotyping approaches do not require labeled examples of cases and controls. Unsupervised approaches identify complex associations between key factors and can quantify them as redefined latent variables. In chapter two I used one type of unsupervised algorithm called tensor decomposition to identify multiple latent sub-phenotypes of preterm birth. Each latent sub-phenotype was distinguished based on its comorbidity pattern and longitudinal disease trajectory. Using the weight assigned to each individual with preterm birth, I showed that the genetic risk for comorbidities are associated with specific sub-phenotypes. A strength of unsupervised approach is the unbiased modeling of high-dimensional datasets. For heterogenous traits, determining the specific disease feature to use to subset individuals is not always obvious. For preterm birth, multiple classification schemes have been proposed and no universal definition exists for stratifying by gestational length. Furthermore, a majority of preterm births present with multiple comorbidities[31] and some increase the risk of prematurity[30]. Thus, analytical approaches such as tensor decomposition are well suited for refining the phenotypic heterogeneity of preterm birth. For future analyses, other variations of tensor decomposition can be applied to similar datasets optimized for specific types of discoveries. More generally, other unsupervised approaches, including machine learning methods such as autoencoders, should be applied. It is possible that each approach may uncover distinct patterns. Additionally, the flexibility of machine learning based models will enable combining diverse data types from EHRs such as clinical labs, medications, and more extensive family and personal histories.

5.3 Evolutionary analyses can lead to insights for disease risk and prioritize candidate regions for functional analyses

5.3.1 Incorporating a holistic evolutionary perspective can yield population specific functional insights

Understanding the human genome, genetic variation, and the genetic basis of diseases requires studying how the genome evolves. Across many traits, robust evidence of strong directional selection has been reported. However, the effect of other types of selection (balancing selection), evolutionary history across different time scales (recent vs. ancient), and how these different modes have jointly acted on the trait associated genomic regions remains poorly understood. In the context of preterm birth, a recent study incorporating multiple measures of positive selection examined the progesterone receptor locus. This locus has experienced recent positive selection in east Asians but remained highly polymorphic in European women. Investigating the conserved variants in this region uncovered associations with spontaneous preterm birth in a cohort of African American women[90]. This study demonstrates that incorporating an evolutionary perspective can lead to new insights for observed disparities in preterm birth risk across populations.

5.3.2 Evolutionary forces prioritize genomic candidates associated with preterm birth for further investigation

Using the largest genome-wide association study, I demonstrated in chapter three that diverse evolutionary forces have acted on regions associated with preterm birth. These include signatures of excess population differentiation, accelerated evolution, and balanced polymorphism. For many of the genomic regions we investigated, the association with preterm birth only reached nominal significance. However, many of these regions had been shaped by evolutionary forces suggestive of functional importance and, when combined with lines of molecular evidence, our results suggest that these genomic regions should evaluated using model organisms to determine any functional consequences. Additionally, evolutionary analyses are efficient compared to large scale genome-wide association studies. There are practical, technical, and economical costs to conducting large scale genome-wide association studies. Evolutionary analyses use existing databases of human variation and methods to detect specific pattern of natural selection. Thus, these analyses can be applied to summary statistics from any genome-wide association study relatively quickly.

5.3.3 Future evolutionary analyses should incorporate effect sizes and cross-trait correlations

There are several future directions for improving the evolutionary analyses for detecting multiple modes of selection from summary statistics of genome-wide association studies. A recurring challenge for genome-wide association studies is properly correcting for population structure. This concern is especially salient as larger studies are conducted on combined cohorts across biobanks. Evidence for selection on height based on effect sizes derived from summary statistics have been shown to be inflated due to population stratification[236,237]. Since the evolutionary analyses I performed do not use effect sizes, this approach is likely unaffected by biases from population stratification. Nevertheless, effect sizes at individual variants may provide another important variable to incorporate into polygenic models of selection. Thus, new methods to correct for population stratification induced inflation in effect sizes will be required before incorporating effect sizes into polygenic models of selection.

Another key direction for improving evolutionary analyses is incorporating cross-trait correlations. The rapid success and increase in the number of genome-wide association studies has expanded the number of unique traits mapped to their associated genomic regions. Studies of evolutionary forces on traits tend to focus on a single trait of interest. However, there is evidence of complex evolutionary dynamics such as trade-offs[94,96,238] or antagonistic selection[239]. New statistical frameworks are being developed to detect multi-trait evolutionary effects[240]. Further development of methods that incorporate phenotypic correlations will expand our understanding of how natural selection shapes the human genome.

5.4 Machine learning can improve pregnancy outcomes but requires rigorous validation in diverse populations

5.4.1 Machine learning applications in pregnancy lags behind other clinical domains

The progress and the pace of machine learning, and more generally artificial intelligence, approaches to biomedical settings has been remarkable. Early success in machine learning was driven by advances in image processing and computer vision. Thus, clinical domains such as a pathology, cardiology, and dermatology that rely heavily on medical imaging have made substantial progress in applying machine learning to predictive tasks[218,241]. Obstetrics has remained slower to adopt machine learning approaches[188]. In pregnancy, patients are frequently monitored and those at high-risk for adverse outcomes receive even more clinical surveillance. Additionally, many outcomes in pregnancy are well defined and can be mapped to common gestational length timeframe.

As a proof-of-concept, I demonstrated how integrating diverse data types from EHRs coupled with genetic data can predict preterm birth in chapter four. I evaluated the performance of this model with different datatypes, clinical contexts, and external datasets. The robust performance of the prediction models can be improved even more with larger datasets and more powerful machine learning models. While EHRs combined with machine learning models holds potential for improving maternal health[242], future studies should rigorously evaluate models for clinical utility and incorporate diverse datasets while developing approaches to mitigate societal biases.

5.4.2 The measure of a model: evaluating utility is necessary for clinical adoption

When evaluating how well a model performs, the area under the receiver operating curve (ROC) is commonly used, in part because it concisely quantifies a model's discriminative ability. In many clinical settings, the target we aim to predict or diagnose is uncommon, if not rare. Machine learning approaches can report a skewed accuracy by performing well on the majority class, but poorly on the rarer, minority class. Thus, using measures of accuracy beyond the ROC analysis is critical. In chapter four, I evaluated the preterm birth prediction model using both ROC and precision-recall curves, which take into account the disease prevalence[208,243]. Future models can be optimized to quantify disease risk. In addition to a binary prediction, quantifying disease risk can complement a clinician's expertise and aid in decision making[244]. Effective risk stratification will require accurate predicted risk probabilities that are evaluated using multiple calibration metrics[245].

Metrics evaluating model performance on training and validation datasets do not directly measure whether a model will be clinically useful, an endpoint that is critical for the ultimate implementation of these tools[222]. Measuring clinical utility must take use a holistic perspective that considers the many facets of health care delivery[222]. For example, is the model output informative in selecting effective clinical interventions? Additionally, the risks and benefits of each intervention to the patient must also be considered. Cost, treatment duration, and adherence are also important factors that shape patient outcomes. Early studies of machine learning applications used historical data from EHRs. The promising findings from these retrospective studies must be followed up with prospective studies[246,247] that evaluate both reproducibility and clinical utility. Guidelines for such prospective trials, including randomized control trials, are now being considered[248] and will likely evolve as this technology is adopted into clinical practice.

Prospective trials highlight the importance of developing models that are robust across time[246,249]. After a model has been validated, it can be deployed relatively quickly into a hospital environment. In addition to an already dynamic healthcare system reacting to societal, economic, and political changes, integrating a predictive model will also modify the baseline characteristics of the patient populations. Better risk reducing therapies (e.g., lipid lowering agents) will improve outcomes and prioritize new risk factors for clinical intervention. Thus, machine learning models must have built-in systems to adapt as the original data it was trained on become outdated [218,241,250]. Careful development and rigorous validation will enable machine learning systems to realize the vision of precision medicine[251]

For preterm birth prediction, future studies should use a prospective study design with multiple endpoints and clinician feedback. Predictive models should output both binary predictions as well as quantify risk for preterm birth in real time during a pregnancy. Machine learning models should also aim to generate personalized screening guidelines for women according to the past medical history and genetic risk. Likewise, although there are only limited interventions for preterm birth, machine learning approaches may be able to predict which women could benefit the most for existing therapies.

5.4.3 Future machine learning models must identify and mitigate societal biases

While developing accurate and reproducible machine learning models, identifying and mitigating model bias remains challenging[223,246,252]. Almost all machine learning models require large datasets to learn from. Since generating new datasets are time consuming and expensive, many models rely on existing datasets from large electronic health record databases. However, these databases have ingrained in themselves many of the biases in society[223,253]. For example, hospital mortality algorithms have varying accuracy by ethnicity[254]. Even more apparent are models classifying skin lesions as benign or malignant that underperform in individuals with darker skin tones[255,256]. Missing data are another feature that can impact model performance. Since interventions are not uniform across patients and rely on specific indications, the act of performing an intervention (such as a laboratory test) maybe more informative than the result of that intervention[257].

Reducing bias in artificial intelligence models will require multiple approaches[257]. To generate equitable models that serve all individuals regardless of socio-demographic factors, the datasets used to train models should be generated and curated thoughtfully. Existing standards such as the PROBAST tool can aid in risk of bias assessment[258]. Another approach leverages the strengths of artificial intelligence to intentionally design algorithms that reduce existing disparities[253]. Unsupervised models that do not require potentially biased labels can identify structure within datasets born out of implicit biases in clinical practice. For example, greater documentation of anxiety and pain was found more often in white patients compared to nonwhite patients[259]. Additionally, transfer learning paradigms can leverage multi-ethnic cohorts to reduce health care disparities[260].

In the context of preterm birth, machine learning models should incorporate socio-economic and demographic factors in model development. We should simultaneously develop new tools to detect and quantify bias in predictive models. Disparity aware models can use different endpoints to illuminate and mitigate existing bias[253]. EHRs also provide an opportunity to redefine healthy ranges along clinically relevant dimensions such as race. For pregnancy, we could define a growth-adjusted gestational age combining fetal and maternal pregnancy data that incorporate race and socio-economic factors. Including diverse datasets from different geographic locations and identifying unique and shared features in each dataset is another way to illuminate differences and leverage them to improve predictive models.

5.4.4 Conclusion

Discovering successful treatments for preterm birth will depend on our understanding of the biological mechanisms underlying birth timing. While broad principles and key molecular pathways of birth timing have been identified[261], the precise mechanisms are unknown and predicting preterm birth remains challenging. In this work, I took a multi-disciplinary approach to examining the phenotypic and genetic heterogeneity of preterm birth. The results of this approach have identified subtypes of preterm birth, prioritized genomic regions for further study, and demonstrated a proof-of-concept of a predictive algorithm using EHRs and genetic data. The complementarity of these different approaches demonstrates the potential for translating biological and genetic discoveries to improve patient care. As genomic and electronic health record databases grow, the tools and methods developed in this work will become even more powerful for understanding the phenotypic and genetic heterogeneity of preterm birth.

6 References

1. Liu, L. *et al.* Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet* **388**, 3027–3035 (2016).

2. Blencowe, H. *et al.* National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet Lond Engl* **379**, 2162–72 (2012).

3. Martin, J. A., Hamilton, B. E., Osterman, M. J., Curtin, S. C. & Matthews, T. J. Births: final data for 2013. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* **64**, 1–65 (2015).

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

5. Purisch, S. E. & Gyamfi-Bannerman, C. Epidemiology of preterm birth. *Semin Perinatol* **41**, 387–391 (2017).

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

7. Outcomes, I. of M. (US) C. on U. P. B. and A. H., Behrman, R. E. & Butler, A. S. Preterm Birth: Causes, Consequences, and Prevention. (2007) doi:10.17226/11622.

8. Barfield, W. D. Public Health Implications of Very Preterm Birth. *Clin Perinatol* **45**, 565–577 (2018).

9. Callaghan, W. M., MacDorman, M. F., Shapiro-Mendoza, C. K. & Barfield, W. D. Explaining the Recent Decrease in US Infant Mortality Rate, 2007–2013. *Obstet Gynecol Surv* **72**, 266–268 (2017).

10. Patel, R. Short- and Long-Term Outcomes for Extremely Preterm Infants. *Am J Perinat* **33**, 318–328 (2016).

11. Mattison, D. R., Wilson, S., Coussens, C. & Gilbert, D. The Role of Environmental Hazards in Premature Birth. https://www.ncbi.nlm.nih.gov/books/NBK216221/#ddd00021 (2003).

12. Platt, M. J. Outcomes in preterm infants. *Public Health* **128**, 399–403 (2014).

13. Marlow, N., Wolke, D., Bracewell, M. A., Samara, M. & Group, Epic. S. Neurologic and Developmental Disability at Six Years of Age after Extremely Preterm Birth. *New Engl J Medicine* **352**, 9–19 (2005).

14. Henderson, J., Carson, C. & Redshaw, M. Impact of preterm birth on maternal well-being and women's perceptions of their baby: a population-based survey. *Bmj Open* **6**, e012676 (2016).

15. Reddy, U. M. *et al.* Serious maternal complications after early preterm delivery (24-33 weeks' gestation). *Am J Obstet Gynecol* **213**, 538.e1–9 (2015).

16. Thanh, B. Y. L. *et al.* Mode of delivery and pregnancy outcomes in preterm birth: a secondary analysis of the WHO Global and Multi-country Surveys. *Sci Rep-uk* **9**, 15556 (2019).

17. Behrman, R. E., Butler, A. S. & Outcomes, C. on U. P. B. and A. H. Preterm Birth: Causes, Consequences, and Prevention. (2007).

18. Dunlop, A. L. *et al.* Racial and geographic variation in effects of maternal education and neighborhood-level measures of socioeconomic status on gestational age at birth: Findings from the ECHO cohorts. *Plos One* **16**, e0245064 (2021).

19. Manuck, T. A. Racial and ethnic differences in preterm birth: A complex, multifactorial problem. *Semin Perinatol* **41**, 511–518 (2017).

20. Stamilio, D. M., Gross, G. A., Shanks, A., DeFranco, E. & Chang, J. J. Racial disparity in preterm birth: a study by Kistka et al. *Am J Obstet Gynecol* **196**, 189–190 (2007).

21. Smith, L. K., Draper, E. S., Manktelow, B. N., Dorling, J. S. & Field, D. J. Socioeconomic inequalities in very preterm birth rates. *Archives Dis Child - Fetal Neonatal Ed* **92**, F11 (2007).

22. Brett, K. M., Strogatz, D. S. & Savitz, D. A. Employment, job strain, and preterm delivery among women in North Carolina. *Am J Public Health* **87**, 199–204 (1997).

23. Janevic, T. *et al.* Neighborhood Deprivation and Adverse Birth Outcomes among Diverse Ethnic Groups. *Ann Epidemiol* **20**, 445–451 (2010).

24. Schempf, A. H., Kaufman, J. S., Messer, L. C. & Mendola, P. The Neighborhood Contribution to Black-White Perinatal Disparities: An Example From Two North Carolina Counties, 1999–2001. *Am J Epidemiol* **174**, 744–752 (2011).

25. Adams, M. M. *et al.* Preterm delivery among black and white enlisted women in the United States Army. *Obstet Gynecol* **81**, 65–71 (1993).

26. Hendler, I. *et al.* The Preterm Prediction study: Association between maternal body mass index and spontaneous and indicated preterm birth. *Am J Obstet Gynecol* **192**, 882–886 (2005).

27. Mercer, B. M. *et al.* The Preterm Prediction Study: Effect of gestational age and cause of preterm birth on subsequent obstetric outcome. *Am J Obstet Gynecol* **181**, 1216–1221 (1999).

28. Conde-Agudelo, A., Rosas-Bermúdez, A. & Kafury-Goeta, A. C. Birth Spacing and Risk of Adverse Perinatal Outcomes: A Meta-analysis. *Jama* **295**, 1809 (2006).

29. Krupa, F. G., Faltin, D., Cecatti, J. G., Surita, F. G. C. & Souza, J. P. Predictors of preterm birth. *Int J Gynecol Amp Obstetrics* **94**, 5–11 (2006).

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

31. Barros, F. C. *et al.* The Distribution of Clinical Phenotypes of Preterm Birth Syndrome. *JAMA Pediatrics* **169**, 220–10 (2015).

32. Epstein, F. H., Goldenberg, R. L., Hauth, J. C. & Andrews, W. W. Intrauterine Infection and Preterm Delivery. *New Engl J Medicine* **342**, 1500–1507 (2000).

33. Mueller-Heubach, E., Rubinstein, D. N. & Schwarz, S. S. Histologic chorioamnionitis and preterm delivery in different patient populations. *Obstet Gynecol* **75**, 622–6 (1990).

34. Andrews, W. W., Goldenberg, R. L. & Hauth, J. C. Preterm labor: emerging role of genital tract infections. *Infect Agents Dis* **4**, 196–211 (1995).

35. Andres, R. L. & Day, M.-C. Perinatal complications associated with maternal tobacco use. *Seminars Neonatol* **5**, 231–241 (2000).

36. Cnattingius, S. The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine Tob Res* **6**, S125–S140 (2004).

37. Gouin, K., Murphy, K., Shah, P. S. & Births, K. S. group on D. of L. B. W. and P. Effects of cocaine use during pregnancy on low birthweight and preterm birth: systematic review and metaanalyses. *Am J Obstet Gynecol* **204**, 340.e1-340.e12 (2011).

38. Bennett, A. D. Perinatal substance abuse and the drug-exposed neonate. *Adv Nurse Pract* **7**, 32–6; quiz 37–8 (1999).

39. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *Bjog Int J Obstetrics Gynaecol* **107**, 375–381 (2000).

40. Kistka, Z. A. F. *et al.* Heritability of parturition timing: an extended twin design analysis. *American Journal of Obstetrics and Gynecology* **199**, 43.e1-43.e5 (2008).

41. III, J. F. S. *et al.* Spontaneous preterm birth: advances toward the discovery of genetic predisposition. *American Journal of Obstetrics and Gynecology* **218**, 294-314.e2 (2018).

42. Chittoor, G. *et al.* Localization of a major susceptibility locus influencing preterm birth. *Mol Hum Reprod* **19**, 687–696 (2013).

43. Karjalainen, M. K. *et al.* A Potential Novel Spontaneous Preterm Birth Gene, AR, Identified by Linkage and Association Analysis of X Chromosomal Markers. *Plos One* **7**, e51378 (2012).

44. Monangi, N. K., Brockway, H. M., House, M., Zhang, G. & Muglia, L. J. The genetics of preterm birth: Progress and promise. *Semin Perinatol* **39**, 574–583 (2015).

45. Wu, W. *et al.* A Genome-Wide Association Study of spontaneous preterm birth in a European population. *F1000research* **2**, 255 (2013).

46. Zhang, H. *et al.* A Genome-Wide Association Study of Early Spontaneous Preterm Delivery. *Genet Epidemiol* **39**, 217–226 (2015).

47. Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *New Engl J Medicine* **377**, 1156–1167 (2017).

48. Knijnenburg, T. A. *et al.* Genomic and molecular characterization of preterm birth. *P Natl Acad Sci Usa* **116**, 5819–5827 (2019).

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

50. Hong, X. *et al.* Genome-wide approach identifies a novel gene-maternal pre-pregnancy BMI interaction on preterm birth. *Nature communications* **8**, 15608 (2017).

51. Pereyra, S., Bertoni, B. & Sapiro, R. Interactions between environmental factors and maternal–fetal genetic variations: strategies to elucidate risks of preterm birth. *European Journal of Obstetrics and Gynecology* **202**, 20–25 (2016).

52. Romero, R. *et al.* The role of infection in preterm labour and delivery. *Paediatr Perinat Ep* **15**, 41–56 (2001).

53. Macones, G. A. *et al.* A polymorphism in the promoter region of TNF and bacterial vaginosis: preliminary evidence of gene-environment interaction in the etiology of spontaneous preterm birth. *Am J Obstet Gynecol* **190**, 1504–1508 (2004).

54. Cha, J. *et al.* Combinatory approaches prevent preterm birth profoundly exacerbated by gene-environment interactions. *J Clin Invest* **123**, 4063–4075 (2013).

55. Kim, Y. M. *et al.* Failure of physiologic transformation of the spiral arteries in patients with preterm labor and intact membranes. *Am J Obstet Gynecol* **189**, 1063–1069 (2003).

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

57. Henderson, J. J., McWilliam, O. A., Newnham, J. P. & Pennell, C. E. Preterm birth aetiology 2004-2008. Maternal factors associated with three phenotypes: spontaneous preterm labour,

preterm pre-labour rupture of membranes and medically indicated preterm birth. *J Maternal-fetal Neonatal Medicine Official J European Assoc Périnat Medicine Fed Asia Ocean Périnat Soc Int Soc Périnat Obstetricians* **25**, 642–7 (2011).

58. Manuck, T. A. *et al.* The phenotype of spontaneous preterm birth: application of a clinical phenotyping tool. *Am J Obstet Gynecol* **212**, 487.e1-487.e11 (2015).

59. Esplin, M. S. *et al.* Cluster analysis of spontaneous preterm birth phenotypes identifies potential associations among preterm birth mechanisms. *Am J Obstet Gynecol* **213**, 429.e1–9 (2015).

60. Stout, M. J., Busam, R., Macones, G. A. & Tuuli, M. G. Spontaneous and indicated preterm birth subtypes: interobserver agreement and accuracy of classification. - PubMed - NCBI. *American Journal of Obstetrics and Gynecology* **211**, 530.e1-530.e4 (2014).

61. Esplin, M. S. The Importance of Clinical Phenotype in Understanding and Preventing Spontaneous Preterm Birth. *American Journal of Perinatology* **33**, 236–244 (2016).

62. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* **15**, 761–771 (2013).

63. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *dx.doi.org* **17**, 353–373 (2016).

64. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *Plos Med* **12**, e1001779 (2015).

65. Erlebacher, A. & Fisher, S. J. Baby's First Organ. *Nature Publishing Group* **317**, 46–53 (2017).

66. Abbot, P. & Capra, J. A. Reproduction: What is a placental mammal anyway? **6**, e30994 (2017).

67. Rawn, S. M. & Cross, J. C. The Evolution, Regulation, and Function of Placenta-Specific Genes. **24**, 159–181 (2008).

68. Roberts, R. M., Green, J. A. & Schulz, L. C. The evolution of the placenta. *Reproduction* **152**, R179-89 (2016).

69. Elliot, M. G. & Crespi, B. J. Phylogenetic Evidence for Early Hemochorial Placentation in Eutheria. *Placenta* **30**, 949–967 (2009).

70. Guo, J. *et al.* Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat Commun* **9**, 1865 (2018).

71. Carter, J. *et al.* Development and validation of predictive models for QUiPP App v.2: tool for predicting preterm birth in women with symptoms of threatened preterm labor. *Ultrasound Obst Gyn* **55**, 357–367 (2020).

72. Son, E. S. M. M. Predicting preterm birth: Cervical length and fetal fibronectin. *Seminars in Perinatology* **41**, 445–451 (2017).

73. Koullali, B., Oudijk, M. A., Nijman, T. A. J., Mol, B. W. J. & Pajkrt, E. Risk assessment and management to prevent preterm birth. *Seminars Fetal Neonatal Medicine* **21**, 80–88 (2016).

74. Kistka, Z. A. F. *et al.* Racial disparity in the frequency of recurrence of preterm birth. - PubMed - NCBI. *American Journal of Obstetrics and Gynecology* **196**, 131.e1-131.e6 (2007).

75. Vink, J. & Feltovich, H. Cervical etiology of spontaneous preterm birth. *Seminars Fetal Neonatal Medicine* **21**, 106–112 (2016).

76. Prinzbach, A. *et al.* Comorbidities in Childhood Celiac Disease: A Phenome Wide Association Study using the Electronic Health Record. *Journal of Pediatric Gastroenterology and Nutrition* **Publish Ahead of Print**, 1 (2018).

77. Zhao, J. *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Scientific Reports* **9**, 1–10 (2019).

78. Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hgg Adv* **2**, 100017 (2020).

79. Abraham, A. *et al.* Dense phenotyping from EHRs enables machine-learning-based prediction of preterm birth. (n.d.) doi:10.1101/2020.07.15.20154864.

80. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

81. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* **86**, 560–72 (2010).

82. Stevens, K., Kegelmeyer, P., Andrzejewski, D. & Buttler, D. Exploring Topic Coherence over Many Models and Many Topics. https://www.aclweb.org/anthology/D12-1087.pdf (2012).

83. Zhao, J. *et al.* Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study. *J Biomed Inform* **98**, 103270 (2019).

84. Lambert, S. A. *et al.* The Polygenic Score Catalog: an open database for reproducibility and systematic evaluation. (n.d.) doi:10.1101/2020.05.20.20108217.

85. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

86. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5–22 (2017).

87. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting Natural Selection in Genomic Data. *Annu Rev Genet* **47**, 97–120 (2013).

88. Siewert, K. M. & Voight, B. F. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Mol Biol Evol* **34**, 2996–3005 (2017).

89. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *Plos Genet* **10**, e1004342 (2014).

90. Li, J. *et al.* Natural Selection Has Differentiated the Progesterone Receptor among Human Populations. *Am J Hum Genetics* **103**, 45–57 (2018).

91. Zeng, J. *et al.* Bayesian analysis of GWAS summary data reveals differential signatures of natural selection across human complex traits and functional genomic categories. *Biorxiv* 752527 (2019) doi:10.1101/752527.

92. Eidem, H. R., McGary, K. L., Capra, J. A., Abbot, P. & Rokas, A. The transformative potential of an integrative approach to pregnancy. *Placenta* **57**, 204–215 (2017).

93. Abbot, P. & Rokas, A. Mammalian pregnancy. *Curr Biol* **27**, R127–R128 (2017).

94. Moon, J. M., Capra, J. A., Abbot, P. & Rokas, A. Immune Regulation in Eutherian Pregnancy: Live Birth Coevolved with Novel Immune Genes and Gene Regulation. *Bioessays* **41**, 1900072 (2019).

95. Rosenberg, K. & Trevathan, W. Bipedalism and human birth: The obstetrical dilemma revisited. *Evol Anthropology Issues News Rev* **4**, 161–168 (1995).

96. Pavličev, M., Romero, R. & Mitteroecker, P. Evolution of the human pelvis and obstructed labor: New explanations of an old obstetrical dilemma. *Am J Obstet Gynecol* **222**, 3–16 (2019).

97. Krogman, W. M. The Scars of Human Evolution. *Sci Am* **185**, 54–57 (1951).

98. Martin, J. A., Hamilton, B. E. & Osterman, and M. J. K. Births in the United States, 2016. https://www.cdc.gov/nchs/data/databriefs/db287.pdf (2017).

99. ESPLIN, M. S. Overview of Spontaneous Preterm Birth: A Complex and Multifactorial Phenotype. *Clin Obstetrics Gynecol* **57**, 518–530 (2014).

100. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med* **5**, 34 (2013).

101. Plunkett, J. *et al.* Mother's genome or maternally-inherited genes acting in the fetus influence gestational age in familial preterm birth. *Hum Hered* **68**, 209–19 (2009).

102. Kjeldbjerg, A. L., Villesen, P., Aagaard, L. & Pedersen, F. S. Gene conversion and purifying selection of a placenta-specific ERV-V envelope gene during simian evolution. *Bmc Evol Biol* **8**, 266 (2008).

103. Phillips, J. B., Abbot, P. & Rokas, A. Is preterm birth a human-specific syndrome? *Evol Medicine Public Heal* **2015**, 136–48 (2015).

104. Chen, C. *et al.* The human progesterone receptor shows evidence of adaptive evolution associated with its ability to act as a transcription factor. *Mol Phylogenet Evol* **47**, 637–649 (2008).

105. Newnham, J. P. *et al.* Strategies to Prevent Preterm Birth. *Front Immunol* **5**, 584 (2014).

106. Akey, J. M. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* **19**, 711–722 (2009).

107. Pybus, M. *et al.* 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res* **42**, D903-9 (2013).

108. Stern, A. J. & Nielsen, R. Handbook of Statistical Genomics. 397–40 (2019) doi:10.1002/9781119487845.ch14.

109. Booker, T. R., Jackson, B. C. & Keightley, P. D. Detecting positive selection in the genome. *Bmc Biol* **15**, 98 (2017).

110. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am J Hum Genetics* **105**, 456–476 (2019).

111. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* **50**, 746–753 (2018).

112. Guo, J., Yang, J. & Visscher, P. M. Leveraging GWAS for complex traits to detect signatures of natural selection in humans. *Curr Opin Genet Dev* **53**, 9–14 (2018).

113. Plunkett, J. *et al.* An Evolutionary Genomic Approach to Identify Genes Involved in Human Birth Timing. *Plos Genet* **7**, e1001365 (2011).

114. Gu, T.-P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606–610 (2011).

115. Tsukada, Y., Akiyama, T. & Nakayama, K. I. Maternal TET3 is dispensable for embryonic development but is required for neonatal growth. *Sci Rep-uk* **5**, 15876 (2015).

116. Liong, S., Quinzio, M. K. W. D., Fleming, G., Permezel, M. & Georgiou, H. M. Is vitamin D binding protein a novel predictor of labour? *Plos One* **8**, e76490 (2013).

117. Sõber, S. *et al.* Extensive shift in placental transcriptome profile in preeclampsia and placental origin of adverse pregnancy outcomes. *Sci Rep-uk* **5**, 13336 (2015).

118. Fitzgerald, E., Boardman, J. P. & Drake, A. J. Preterm Birth and the Risk of Neurodevelopmental Disorders - Is There a Role for Epigenetic Dysregulation? *Curr Genomics* **19**, 507–521 (2018).

119. Zelko, I. N., Zhu, J. & Roman, J. Maternal undernutrition during pregnancy alters the epigenetic landscape and the expression of endothelial function genes in male progeny. *Nutr Res* **61**, 53–63 (2019).

120. Paule, S. G., Airey, L. M., Li, Y., Stephens, A. N. & Nie, G. Proteomic Approach Identifies Alterations in Cytoskeletal Remodelling Proteins during Decidualization of Human Endometrial Stromal Cells. *J Proteome Res* **9**, 5739–5747 (2010).

121. Sitras, V. *et al.* Differential Placental Gene Expression in Severe Preeclampsia. *Placenta* **30**, 424–433 (2009).

122. Meunier, J.-C. *et al.* Isolation and structure of the endogenous agonist of opioid receptor-like ORL1 receptor. *Nature* **377**, 532–535 (1995).

123. BH, D. Uterus-Relaxing Effects of Nociceptin and Nocistatin: Studies on Preterm and Term-Pregnant Human Myometrium In vitro. *Reproductive Syst Sex Disord* **02**, (2013).

124. Gáspár, R., Deák, B. H., Klukovits, A., Ducza, E. & Tekes, K. Nociceptin Opioid. *Vitamins Hormones* **97**, 223–240 (2015).

125. JUNG, K.-H. *et al.* Associations of Vitamin D Binding Protein Gene Polymorphisms with the Development of Peripheral Arthritis and Uveitis in Ankylosing Spondylitis. *J Rheumatology* **38**, 2224–2229 (2011).

126. Muindi, J. R. *et al.* Serum Vitamin D Metabolites in Colorectal Cancer Patients Receiving Cholecalciferol Supplementation: Correlation with Polymorphisms in the Vitamin D Genes. *Hormones Cancer* **4**, 242–250 (2013).

127. Zhou, S.-S., Tao, Y.-H., Huang, K., Zhu, B.-B. & Tao, F.-B. Vitamin D and risk of preterm birth: Up-to-date meta-analysis of randomized controlled trials and observational studies. *J Obstetrics Gynaecol Res* **43**, 247–256 (2017).

128. D'Silva, A. M., Hyett, J. A. & Coorssen, J. R. Proteomic analysis of first trimester maternal serum to identify candidate biomarkers potentially predictive of spontaneous preterm birth. *J Proteomics* **178**, 31–42 (2018).

129. Burris, H. H. *et al.* Plasma 25-hydroxyvitamin D during pregnancy and small-for-gestational age in black and white infants. *Ann Epidemiol* **22**, 581–586 (2012).

130. Jablonski, N. G. & Chaplin, G. The roles of vitamin D and cutaneous vitamin D production in human evolution and health. *Int J Paleopathol* **23**, 54–59 (2018).

131. Hollis, B. W. & Wagner, C. L. New insights into the vitamin D requirements during pregnancy. *Bone Res* **5**, 17030 (2017).

132. Yang, J. T., Rayburn, H. & Hynes, R. O. Cell adhesion events mediated by alpha 4 integrins are essential in placental and cardiac development. *Dev Camb Engl* **121**, 549–60 (1995).

133. Burrows, T. D., King, A. & Loke, Y. W. Trophoblast migration during human placental implantation. *Placenta* **17**, A48 (1996).

134. Mincheva-Nilsson, L. & Baranov, V. The Role of Placental Exosomes in Reproduction: PLACENTAL EXOSOMES IN REPRODUCTION. *Am J Reprod Immunol* **63**, 520–533 (2010).

135. Paidas, M. J. *et al.* A genomic and proteomic investigation of the impact of preimplantation factor on human decidual cells. *Am J Obstet Gynecol* **202**, 459.e1-459.e8 (2010).

136. Paule, S., Li, Y. & Nie, G. Cytoskeletal remodelling proteins identified in fetal-maternal interface in pregnant women and rhesus monkeys. *J Mol Histol* **42**, 161–166 (2011).

137. Strohl, A. *et al.* Decreased adherence and spontaneous separation of fetal membrane layers--amnion and choriodecidua--a possible part of the normal weakening process. *Placenta* **31**, 18–24 (2009).

138. Plunkett, J. *et al.* Primate-specific evolution of noncoding element insertion into PLA2G4C and human preterm birth. *Bmc Med Genomics* **3**, 62 (2010).

139. Rosenberg, K. & Trevathan, W. Birth, obstetrics and human evolution. *Bjog Int J Obstetrics Gynaecol* **109**, 1199–1206 (2002).

140. Srinivasan, S. *et al.* Genetic Markers of Human Evolution Are Enriched in Schizophrenia. *Biol Psychiat* **80**, 284–92 (2015).

141. Sun, S.-C. *et al.* Actin nucleator Arp2/3 complex is essential for mouse preimplantation embryo development. *Reproduction Fertility Dev* **25**, 617–623 (2012).

142. Majewska, M. *et al.* Placenta Transcriptome Profiling in Intrauterine Growth Restriction (IUGR). *Int J Mol Sci* **20**, 1510 (2019).

143. Ferrer-Admetlla, A. *et al.* Balancing Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes. *J Immunol* **181**, 1315–1322 (2008).

144. Siewert, K. M., evolution, B. V. M. biology and & 2017. Detecting long-term balancing selection using allele frequency correlation. *academic.oup.com* (n.d.).

145. Mor, G. & Cardenas, I. The Immune System in Pregnancy: A Unique Complexity: IMMUNE SYSTEM IN PREGNANCY. *Am J Reprod Immunol* **63**, 425–433 (2010).

146. York, T. P., Eaves, L. J., Neale, M. C. & Strauss, J. F. The contribution of genetic and environmental factors to the duration of pregnancy. *Am J Obstet Gynecol* **210**, 398–405 (2014).

147. Manuck, T. A. *et al.* Admixture Mapping to Identify Spontaneous Preterm Birth Susceptibility Loci in African Americans. *Obstetrics Gynecol* **117**, 1078–1084 (2011).

148. Zhang, G. *et al.* Genetic Associations With Gestational Duration and Spontaneous Preterm Birth. *Obstet Gynecol Surv* **73**, 83–85 (2018).

149. Byars, S. G. *et al.* Genetic loci associated with coronary artery disease harbor evidence of selection and antagonistic pleiotropy. *Plos Genet* **13**, e1006328 (2017).

150. Muglia, L. J. & Katz, M. The Enigma of Spontaneous Preterm Birth. *Obstetric Anesthesia Dig* **31**, 75–76 (2011).

151. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinform Oxf Engl* **31**, 418–20 (2015).

152. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

153. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinform Oxf Engl* **27**, 2156–8 (2011).

154. Gautier, M., Klassmann, A. & Vitalis, R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular ecology resources* **17**, 78–90 (2017).

155. Siewert, K. M. & Voight, B. F. Detecting Long-term Balancing Selection using Allele Frequency Correlation. *Biorxiv* 112870 (2017) doi:10.1101/112870.

156. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493–D496 (2004).

157. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res* **12**, 996–1006 (2002).

158. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* **49**, 618–624 (2017).

159. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–121 (2010).

160. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901–913 (2005).

161. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790–1797 (2012).

162. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

163. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495–501 (2010).

164. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).

165. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754–D761 (2017).

166. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

167. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184–1191 (2009).

168. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

169. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**, 1812–1819 (2017).

170. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2010).

171. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37–48 (1999).

172. Leigh, J. W. & Bryant, D. popart : full-feature software for haplotype network construction. *Methods Ecol Evol* **6**, 1110–1116 (2015).

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

174. Iams, J. *et al.* The Length of the Cervix and the Risk of Spontaneous Premature Delivery. *New Engl J Medicine* **334**, 567–573 (1996).

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

176. Mazaki-Tovi, S. *et al.* Recurrent Preterm Birth. *Semin Perinatol* **31**, 142–158 (2007).

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

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

179. Francesca, L. *et al.* Biomarkers for predicting spontaneous preterm birth: an umbrella systematic review. *The Journal of Maternal-Fetal & Neonatal Medicine* **0**, 726–734 (2019).

180. Dabi, Y. *et al.* Clinical validation of a model predicting the risk of preterm delivery. *PloS one* **12**, e0171801 (2017).

181. Ngo, T. T. M. *et al.* Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**, 1133–1136 (2018).

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

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

184. Weber, A. *et al.* Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. *Annals of epidemiology* **28**, 783-789.e1 (2018).

185. Baer, R. J. *et al.* Pre-pregnancy or first-trimester risk scoring to identify women at high risk of preterm birth. *European Journal of Obstetrics and Gynecology* **231**, 235–240 (2018).

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

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

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

189. Artzi, N. S. *et al.* Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* **26**, 71–76 (2020).

190. Ravizza, S. *et al.* Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat Med* **25**, 57–59 (2019).

191. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nature Publishing Group* **31**, 1–10 (2020).

192. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).

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

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

195. Krishnapuram, B. *et al.* XGBoost: A Scalable Tree Boosting System. *Arxiv* 785–794 (2016) doi:10.1145/2939672.2939785.

196. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. (2009) doi:10.1007/978-0-387-84858-7.

197. Corey, K. M. *et al.* Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *Plos Med* **15**, e1002701 (2018).

198. Jing, L. *et al.* A Machine Learning Approach to Management of Heart Failure Populations. *Jacc Hear Fail* **8**, 578–587 (2020).

199. Vogel, J. P. *et al.* The global epidemiology of preterm birth. *Best Pract Res Cl Ob* **52**, 3–12 (2018).

200. Smith, G. C. S. & Pell, J. P. Teenage Pregnancy and Risk of Adverse Perinatal Outcomes Associated With First and Second Births: Population Based Retrospective Cohort Study. *Obstet Gynecol Surv* **57**, 136–137 (2002).

201. Waldenström, U. *et al.* Adverse Pregnancy Outcomes Related to Advanced Maternal Age Compared With Smoking and Being Overweight. *Obstetrics Gynecol* **123**, 104–112 (2014).

202. Carolan, M. Maternal age ≥45 years and maternal and perinatal outcomes: A review of the evidence. *Midwifery* **29**, 479–489 (2013).

203. Ray, J. G., Vermeulen, M. J., Shapiro, J. L. & Kenshole, A. B. Maternal and neonatal outcomes in pregestational and gestational diabetes mellitus, and the influence of maternal obesity and weight gain: the DEPOSIT study. *Qjm Int J Medicine* **94**, 347–356 (2001).

204. Whiteman, V. *et al.* Impact of sickle cell disease and thalassemias in infants on birth outcomes. *Eur J Obstet Gyn R B* **170**, 324–328 (2013).

205. Umesawa, M. & Kobashi, G. Epidemiology of hypertensive disorders in pregnancy: prevalence, risk factors, predictors and prognosis. *Hypertens Res* **40**, 213–220 (2017).

206. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).

207. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).

208. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. 233–240 (2006) doi:10.1145/1143844.1143874.

209. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. I. Guyon et al.) 4765--4774 (Curran Associates, Inc., 2017).

210. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* **2**, 749–760 (2018).

211. Creanga, A. A. *et al.* Pregnancy-Related Mortality in the United States, 2006–2010. *Obstetrics Gynecol* **125**, 5–12 (2015).

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

213. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. & Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep-uk* **10**, 11981 (2020).

214. Yan, L. *et al.* An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* **2**, 283–288 (2020).

215. Couronné, R., Probst, P. & Boulesteix, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *Bmc Bioinformatics* **19**, 270 (2018).

216. Gao, C. *et al.* Deep learning predicts extreme preterm birth from electronic health records. *J Biomed Inform* **100**, 103334 (2019).

217. Torchin, H. & Ancel, P.-Y. [Epidemiology and risk factors of preterm birth]. *J De Gynecol Obstetrique Et Biologie De La Reproduction* **45**, 1213–1230 (2016).

218. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **25**, 44–56 (2019).

219. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat Med* **25**, 30–36 (2019).

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

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

222. Shah, N. H., Milstein, A. & PhD, S. C. B. Making Machine Learning Models Clinically Useful. *Jama* **322**, 1351–1352 (2019).

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

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

225. Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. in *International conference on machine learning* 115--123 (2013).

226. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825--2830 (2011).

227. Soysal, E. *et al.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assn* **25**, 331–336 (2017).

228. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* **27**, e1608 (2018).

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

230. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).

231. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Arxiv* (2018).

232. Campello, R. J. G. B., Moulavi, D. & Sander, J. Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II. 160–172 (2013) doi:10.1007/978-3-642-37456-2_14.

233. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

234. Gustafson, E., Pacheco, J., Wehbe, F., Silverberg, J. & Thompson, W. A Machine Learning Algorithm for Identifying Atopic Dermatitis in Adults from Electronic Health Records. *IEEE International Conference on Healthcare Informatics . IEEE International Conference on Healthcare Informatics* **2017**, 83–90 (2017).

235. Wei, W.-Q. *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association* **23**, e20–e27 (2016).

236. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, e39725 (2019).

237. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* **8**, e39702 (2019).

238. Byars, S. G. & Inouye, M. Genetic loci associated with coronary artery disease harbor evidence of selection and antagonistic pleiotropy. http://journals.plos.org/plosgenetics/article/file?id=10.1371/journal.pgen.1006328&type=printable (n.d.).

239. Rodríguez, J. A. *et al.* Antagonistic pleiotropy and mutation accumulation influence human senescence and disease. *Nat Ecol Evol* **1**, 0055 (2017).

240. Stern, A. J., Speidel, L., Zaitlen, N. A. & Nielsen, R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am J Hum Genetics* **108**, 219–239 (2021).

241. Shaik, R. Artificial Intelligence in Healthcare. *Indian J Pharm Pract* **12**, 215–216 (2019).

242. Sadovsky, Y. *et al.* Advancing human health in the decade ahead: pregnancy as a key window for discovery. *Am J Obstet Gynecol* **223**, 312–321 (2020).

243. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *Plos One* **10**, e0118432 (2015).

244. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* **20**, e262–e273 (2019).

245. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assn* **27**, 621–633 (2020).

246. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *Bmc Med* **17**, 195 (2019).

247. Beam, A. L., Manrai, A. K. & Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *Jama* **323**, 305–306 (2020).

248. Rivera, S. C. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digital Heal* **2**, e549–e560 (2020).

249. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat Biomed Eng* **2**, 719–731 (2018).

250. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of Calibration Drift in Clinical Prediction Models to Inform Model Updating. *J Biomed Inform* **112**, 103611 (2020).

251. Denny, J. C. & Collins, F. S. Precision medicine in 2030—seven ways to transform healthcare. *Cell* **184**, 1415–1419 (2021).

252. Price, W. N. Big data and black-box medical algorithms. *Sci Transl Med* **10**, eaao5333 (2018).

253. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* **27**, 136–140 (2021).

254. Chen, I., Johansson, F. D. & Sontag, D. Why Is My Classifier Discriminatory? *Arxiv* (2018).

255. Haenssle, H. A., Fink, C., Rosenberger, A. & Uhlmann, L. Reply to the letter to the editor "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists" by H. A. Haenssle et al. *Ann Oncol Official J European Soc Medical Oncol* **30**, 854–857 (2019).

256. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

257. Parikh, R. B., Teeple, S. & Navathe, A. S. Addressing Bias in Artificial Intelligence in Health Care. *Jama* **322**, 2377–2378 (2019).

258. Wolff, R. F. *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* **170**, 51–58 (2019).

259. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *Ama J Ethics* **21**, E167-179 (2019).

260. Gao, Y. & Cui, Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun* **11**, 5131 (2020).

261. Renzo, G. C. D., Tosto, V. & Giardina, I. The biological basis and prevention of preterm birth. *Best Pract Res Cl Ob* **52**, 13–22 (2018).

262. LaBella, A. L. *et al.* Accounting for diverse evolutionary forces reveals mosaic patterns of selection on human preterm birth loci. *Nat Commun* **11**, 3731 (2020).