

Associations with spontaneous and indicated preterm birth in a densely phenotyped EHR cohort

Jean M. Costello PhD^{1,2*} & Hannah Takasuka^{3*}, Jacquelyn Roger BS⁴, Ophelia Yin MD⁵, Alice Tang BS⁶, Tomiko Oskotsky MD^{1,2}, Marina Sirota PhD^{1,2*}, John A. Capra PhD^{1,7*}

¹Bakar Computational Health Sciences Institute, UCSF; ²Department of Pediatrics, UCSF; ³Department of Oral and Craniofacial Sciences, UCSF; ⁴Biological and Medical Informatics, UCSF; ⁵Obstetrics, Gynecology & Reproductive Sciences Department, UCSF; ⁶Bioengineering, UCSF & UC Berkeley; ⁷Department of Epidemiology and Biostatistics, UCSF

*Authors contributed equally

Correspondence to tony@capralab.org, Marina.Sirota@ucsf.edu

Abstract

Background: Preterm birth (PTB) is the leading cause of infant mortality and follows multiple biological pathways, many of which are poorly understood. Some PTBs result from medically indicated labor following complications from hypertension and/or diabetes, while many others are spontaneous with unknown causes. Previously, investigation of potential risk factors has been limited by lack of data on maternal medical history and the difficulty of classifying PTBs as indicated or spontaneous. Here, we leverage electronic health record (EHR) data (patient health information including demographics, diagnoses, and medications) and a supplemental curated pregnancy database to overcome these limitations. Novel associations may provide new insight into the pathophysiology of PTB as well as help identify individuals who would be at risk of PTB.

Methods: We quantified associations between maternal diagnoses and preterm birth using logistic regression controlling for maternal age and socioeconomic factors within a University of California, San Francisco (UCSF), EHR cohort with 10,643 births ($n_{term} = 9692$, $n_{spontaneous_preterm} = 449$, $n_{indicated_preterm} = 418$) and maternal pre-conception diagnosis phenotypes derived from International Classification of Diseases (ICD) 9 and 10 codes.

31 **Results:** Eighteen conditions significantly and robustly (False Discovery Rate (FDR)<0.05)
32 associated with PTBs compared to term. We discovered known (hypertension, diabetes, and
33 chronic kidney disease) and less established (blood, cardiac, gynecological, and liver
34 conditions) associations. Type 1 diabetes was the most significant overall association
35 (adjusted $p = 1.6 \times 10^{-14}$, adjusted OR = 7 (95% CI 5, 12)), and the odds ratios for the
36 significant phenotypes ranged from 3 to 13. We further carried out analysis stratified by
37 spontaneous vs. indicated PTB. No phenotypes significantly associated with spontaneous
38 PTB; however, the results for indicated PTB largely recapitulated the phenotype associations
39 with all PTBs.

40
41 **Conclusions:** Our study underscores the limitations of approaches that combine indicated
42 and spontaneous births together. When combined, significant associations were almost
43 entirely driven by indicated PTBs, although our spontaneous and indicated groups were of a
44 similar size. Investigating the spontaneous population has the potential to reveal new
45 pathways and understanding of the heterogeneity of PTB.

46
47 **Keywords:** spontaneous preterm birth, indicated preterm birth, electronic health records,
48 diagnosis associations

50 **1 Background**

51 Preterm birth (PTB) is the leading cause of infant mortality worldwide (1) and can result
52 in serious acute and long-term health consequences (2,3). There are multiple proposed
53 pathways for preterm birth, but its etiology remains poorly understood (4–7). About two thirds
54 of PTBs in the US are classified as spontaneous preterm while the remaining third are
55 medically indicated (iatrogenic) preterm (8). An indicated preterm birth is typically initiated
56 based on a list of risk factors, which includes preeclampsia, diabetes complications,
57 intrauterine abnormalities, and placental abnormalities (9). Some of these risk factors, such

58 as poorly managed hypertension, may be present prior to pregnancy. Spontaneous preterm
59 birth, by contrast, lacks a defined set of known risk factors, and the pathophysiology behind
60 it remains poorly understood (8).

61

62 Maternal risk factors for indicated preterm birth include older maternal age, heart
63 disease, hypertension, diabetes, tobacco use, previous preterm delivery, and socioeconomic
64 factors (8,10). Zheng et al. studied lifestyle factors, obstetric and fetal complications,
65 maternal diseases, and socioeconomic factor associations with preterm birth in 3,147 cases
66 and controls across 15 Chinese hospitals (11). They measured multiple pregnancies,
67 hypertensive disorders, and obstetric disorders to be the strongest predictors of iatrogenic
68 preterm birth, with socioeconomic risk factors such as maternal education and prenatal care
69 access also significant.

70 Several maternal risk factors for spontaneous preterm birth have been proposed,
71 including prior spontaneous preterm birth, gynecological anatomy variation, short inter-
72 pregnancy interval, and multiple gestations (12). Prior spontaneous preterm birth is the
73 strongest known risk factor. In the United States, racism is a risk factor for spontaneous
74 preterm birth (13), with higher rates among non-Hispanic Black birthing people when
75 compared to white birthing people, including after adjustment for socioeconomic variables
76 (14). Some studies have explored whether gene-gene and/or gene-environment interactions
77 might exist to explain racial disparities, but these studies are limited to cohorts of a few
78 hundred patients (12).

79

80

81 Improved understanding of pathways and clinical factors leading to preterm birth could
82 lead to better interventions to prevent preterm birth, especially spontaneous preterm birth.
83 Investigating pre-pregnancy conditions associated with subsequent PTB has the potential to
84 generate hypotheses about pathways towards PTB. Many large studies of conditions
85 associated with PTB rely on registry data, which provides limited phenotypic information.

86

87 EHR databases provide dense phenotyping including demographics, diagnoses, and
88 medications over time that can provide insights difficult to obtain from other data sources.
89 However, EHR systems may not distinguish between spontaneous and indicated deliveries.
90 Nonetheless, EHR data are particularly well-suited to the study of pregnancy (15). For
91 instance, machine learning models have used EHR data to accurately predict preterm birth
92 on thousands of patients (16). While complex machine learning models have great potential
93 to improve obstetric and gynecological care, novel insights from straightforward methods
94 applied to EHR data could more easily translate to pathway discovery and evidence-based
95 care.

96

97 In this study, we use logistic regression to measure associations between maternal pre-
98 conception diagnoses and different types of preterm birth using University of California, San
99 Francisco EHR data supplemented with a physician-curated database of delivery
100 information. With this approach, we reproduce widely known preterm birth risk factors
101 including older maternal age and major chronic diseases. Moreover, we discover several
102 new preterm birth associations with less-studied conditions such as decreased white blood
103 count. We also demonstrate that all significant associations with PTB are driven by indicated
104 PTBs and that no diagnoses significantly associate with spontaneous preterm birth.

105 **2 Results**

106 **A densely phenotyped preterm birth cohort linked to electronic health records**

107 To identify potential clinical risk factors for PTB, we defined cohorts of preterm and term
108 deliveries based on curated data from the UCSF Perinatal Database (PDB) and linked
109 these to phenotypes from the UCSF electronic health record (EHR) database.
110 The cohort consisted of 10,643 deliveries to 9,399 individuals from 2001 to 2022 (**Figure**
111 **1a**). There were 975 PTBs in the cohort, which we further classified as spontaneous PTBs

112 (n=449) or indicated PTBs (n=418). The remaining 108 PTBs could not be classified. Each
113 of the preterm groups (spontaneous, indicated, all) was compared to term “controls” born at
114 37 weeks or later (n=9671). More details about the cohorts are provided in the Methods
115 section.

116
117 The demographics of the cohort reflected the population of the San Francisco Bay area
118 served by UCSF. Most individuals had more than 12 years of education (84%). A large
119 majority also used private insurance for the delivery (93.2%). The mean maternal age was
120 34.4 years, and maternal age ranged from 14 years to 55 years. There were no significant
121 differences in maternal age between indicated, spontaneous, and term individuals (**Figure**
122 **S2a**; $p_{\text{indicated-term}} = 0.2$, $p_{\text{spontaneous-term}} = 0.1$, $p_{\text{indicated-spontaneous}} = 0.9$, Mann-Whitney U test). The
123 two most represented self-reported racial categories were single-race white (48.3%) and
124 single-race Asian/Pacific Islander (25%) (Table 1).

125
126 For each individual, we identified all phenotypes present in their EHR before conception
127 (**Figure 1b**). We harmonized phenotypes into phecodes, a curated grouping of ICD codes
128 intended to capture clinically meaningful concepts (**Figure S1**). We identified 1322 unique
129 phecodes across the cohort, and individuals had an average of 8 unique phecodes in their
130 record prior to conception (**Figure S2b**). Most diagnoses (86.8%) were rare — occurring in
131 fewer than 1% of patients (**Figure S3a**). Most medical visits with a diagnosis occurred within
132 2 years before conception (**Figure S2b**); over 95% of individuals’ EHR start date was less
133 than 2.5 years before conception (**Figure S2c**), and the maximum EHR length was 21.7
134 years before conception (**Figure S3b**).

135 136 **Diverse pre-conception phenotypes associate with PTB risk**

137 We tested each of the 1322 phecodes present in the cohort for association with preterm
138 vs. term birth using logistic regression with maternal age, maternal education, insurance

139 status, and race as covariates (**Figure 1c**). Of the covariates, maternal education had the
140 highest rate of missingness at 7.9%. Race was missing for 1.7% of the cohort, and
141 insurance classification was missing for 0.4% of the cohort (Table 1). Missing values for
142 maternal education, race, and insurance classification were encoded as a separate
143 categorical variable. We adjusted for multiple testing by controlling the false discovery rate
144 (FDR) at 5% using the Benjamini-Hochberg procedure. Finally, for each significant
145 association we evaluated its robustness by removing a random individual with the diagnosis
146 and recomputing the association (Methods and **Figure 1c**). We repeated this process 50
147 times. An association which was significant in every iteration is considered robust.

148

149 We identified 34 significant and robust preterm birth associations among the 1322
150 diagnoses tested in the logistic regression (**Figure 2**). As expected, the most significant
151 associations were well-established risk factors: type 1 diabetes (adjusted $P = 1.7 \times 10^{-14}$, $OR =$
152 8 (95% CI 4, 12)), essential hypertension (adjusted $P = 2.1 \times 10^{-12}$, $OR = 3.3$ (95% CI 2.5, 4.5)),
153 and type 2 diabetes (adjusted $P = 8.1 \times 10^{-12}$, $OR = 4.6$ (95% CI 3.1, 6.8)). Since the PTB
154 cohort includes both spontaneous and indicated preterm deliveries, these associations likely
155 reflect medical practice as well as potential intrinsic risk.

156

157 After diabetes and hypertension related diagnoses, chronic kidney disease (CKD) was
158 the next strongest preterm birth association (adjusted $P = 1.6 \times 10^{-5}$). Several other renal
159 conditions were also among the significant associations, including a kidney replaced by
160 transplant and other disorders of the kidney and ureters (**Figure 2b**). Associations between
161 preterm birth and CKD have been observed in studies around the world, but the
162 mechanisms and relevance to risk are not well understood (17).

163

164 The remainder of the significant associations included blood disorders, cardiac
165 conditions, pulmonary conditions, liver conditions, electrolyte imbalances, and digestive
166 conditions. To explore the meaning of the unspecific phenotypes “Other disorders of liver”

167 and “Other diseases of lung,” we extracted concepts from clinical notes using ctnames (18).
168 The “Other disorders of liver” (n=55) phenotype represents conditions including liver lesion
169 (n=20), liver cirrhosis (n=15), liver mass (n=15), liver carcinoma (n=14), and fatty liver
170 (n=13). The “Other diseases of lung” (n=38) phenotype represents conditions including lung
171 consolidation (n=11), interstitial lung diseases (n=5), and lung mass (n=4). Odds ratios, P
172 values, and sample sizes for associations between PTB and all 1322 diagnosis phenotypes
173 are in **Figure 2** and **Table S1**.

174

175 **No diagnoses are associated with spontaneous preterm birth**

176 Given the strong associations with preterm birth overall, we next tested for associations
177 between clinical phenotypes and spontaneous preterm birth vs. term and indicated preterm
178 birth vs. term. No diagnoses were significantly associated with spontaneous preterm birth
179 (**Figure 3a**). In contrast, 30 diagnoses significantly and robustly associated with indicated
180 PTBs (**Figure 3b**). The absence of significant associations with spontaneous preterm birth is
181 not due to lower statistical power than for indicated preterm birth given their similar sample
182 size ($n_{spontaneous} = 449$, $n_{indicated} = 418$).

183

184 Of the 30 diagnoses associated with indicated preterm birth, 28 follow similar trends in
185 spontaneous preterm birth, albeit at much lower effect sizes ($r^2=0.41$, *linear regression*,
186 **Figure 3c**). For example, hypertension has an odds ratio of 6 for indicated and 1.5 for
187 spontaneous. The only two diagnoses with different directions of effect are acute laryngitis
188 and tracheitis and congestive heart failure (CHF) not otherwise specified (NOS); both are
189 significant, robust risk factors for indicated preterm birth, but are protective (though not
190 significant) for spontaneous preterm birth.

191 **Phenotypes associated with all PTBs are also found for indicated PTBs.**

192 Of 18 significant and robust diagnoses associated with all PTBs, 17 are also among the
193 30 significant diagnoses associated with indicated PTBs (**Figure 4**). Diabetes, kidney
194 diseases, and hypertension were the main diagnosis categories associated with overall and
195 indicated PTB. The diagnoses associated with only indicated PTB but not the overall PTB
196 cohort were spread across organs including the liver, lung, and heart. “Disease of tricuspid
197 valve” was the only significant and robust association with overall PTB that was not found for
198 indicated PTB ($P_{BH_indicated} = 0.28$ $OR_{indicated} = 7$ [1.2-42], $P_{BH_overall_preterm} = 8 \times 10^{-3}$, $OR_{overall_preterm} =$
199 11 [3-39], *logistic regression*).

200 **3 Discussion**

201 Our study uses the rich phenotype data present in EHRs to generate hypotheses about
202 the connection between diagnoses prior to pregnancy and risk for indicated and
203 spontaneous PTB. We replicated known associations, including hypertension, diabetes, and
204 chronic kidney disease. We also found several associations that warrant further
205 investigation.

206
207 Most importantly, in our stratified analysis, we found no significant risk factors for
208 spontaneous PTB. This underscores the limitations of approaches that combine indicated
209 and spontaneous births together. When combined, significant associations were entirely
210 driven by indicated PTBs, even though the spontaneous and indicated groups were of a
211 similar size. Thus, our understanding of risk factors for spontaneous PTB remains limited.

212
213 The most significant hits of our study replicated well-established risk factors for PTB, with
214 the four most significant being type 1 diabetes, essential hypertension, type 2 diabetes, and
215 hypertensive heart and/or renal disease. These likely reflect clinical practice as they have
216 existing recommendations for preterm delivery (9). Additionally, several significant phecodes

217 relate to kidney function, such as chronic kidney disease, chronic renal failure, and other
218 disorders of the kidney and ureters. Harel et. al propose that pre-pregnancy counseling,
219 increased monitoring of the mother and fetus, and aspirin treatment to prevent preeclampsia
220 would likely improve pregnancy outcomes for mothers with CKD, as indications for delivery
221 are often hypertensive disorders, worsening renal function, fetal growth restriction or
222 abnormal antenatal testing, or worsening maternal morbidity (17).

223

224 The association between decreased white blood cell count and overall PTB or related
225 conditions is not widely agreed upon. Some studies found no association (19) while others
226 did (20,21). The association between PTB and pre-conception lung conditions including lung
227 consolidation, interstitial lung diseases, and lung mass is not well studied. Our association
228 between PTB and pre-conception lung conditions coincides with the more established
229 correlations that preterm birth causes lung conditions in the newborn through adulthood (22)
230 and that preterm birth tends to pass down in families (23). Furthermore, the association
231 between PTB and liver conditions including liver lesion, liver cirrhosis, liver mass, liver
232 carcinoma, and fatty liver has been previously explored (24–26).

233

234 Comparing the results of the spontaneous subgroup analysis with the indicated subgroup
235 analysis, we see that the two produce very distinct results. Further, we see that the indicated
236 subgroup produces results very similar to the main analysis. This pattern is likely explained
237 by the fact that established risk factors are key to clinicians' decision-making when it comes
238 to indicating delivery. As a result, we expect that the indicated group will have higher rates of
239 these known risk factors and be more homogeneous. Our study highlights the fact that
240 hypothesis-generation studies investigating PTB using data sources unable to distinguish
241 between spontaneous and indicated PTB will have shortcomings. Investigating the
242 heterogeneous spontaneous population has the potential to uncover lesser-known
243 associations that may help reveal new pathways and understanding of PTB.

244

245 A major strength of our study is the use of a curated births database. This database not
246 only records whether deliveries were spontaneous or iatrogenic (information which is not
247 present in our EHR data), but also helps avoid outcome misclassification. Prior work has
248 identified outcome misclassification as a concern for such EHR-based association studies
249 (27). Additionally, researchers investigating obstetric data quality in an EHR system found
250 that quality was varied and recommended manual abstraction where possible (28). By using
251 a database reviewed by physicians to define our outcome, we minimize the risk of such
252 misclassification.

253

254 An additional strength of our study is the wide range of phenotypes we were able to
255 investigate. Our study covered more than 1,300 phenotypes across 17 different phecode
256 categories.

257

258 There are a number of limitations in our study. Our data come from a tertiary care center,
259 which presents two limitations: the patients and deliveries seen at this facility are not
260 representative of the overall local population, and many patients being seen for delivery do
261 not have a previous clinical record at the facility. We present both patient demographics
262 (**Table 1**) and sample selection (**Figure 1**) to show the context in which the study was
263 performed. Compared to the state of California, the PTB rate among our cohort is similar;
264 however, our study population was generally older, more educated, and was privately
265 insured at a high rate. The absence of a significant difference in age distribution between
266 our indicated, spontaneous, and term cohorts may not reproduce in a US age-representative
267 dataset (**Figure S2 a**). We do not have complete medical histories for the individuals
268 included in our study; conditions or deliveries that occurred elsewhere may not be noted in
269 our records.

270

271 Major chronic health conditions appear with high sample sizes and strong statistical
272 power in our cohort. While we captured diagnosis phenotypes across all medical specialties

273 using EHR, diagnosis phenotypes that are not major chronic health conditions will likely be
274 underdiagnosed and under recorded. Consequently, we expected and found a lower sample
275 size and weaker statistical power in our cohort for these phenotypes. More specifically, there
276 are over 585 diagnoses that occur in fewer than 10 individuals (**Figure S3 a**), and over 300
277 of these are $n_{\text{preterm}}=0$.

278

279 We expected that most diagnoses would be recorded in the short time before
280 conception, and we found that over 50% of diagnoses occurred within 1 year of conception
281 (**Figure S3 b**). This represents a common limitation of electronic health record trajectory
282 analysis research, especially prevalent at UCSF as a tertiary care center: patients often use
283 many healthcare institutions over their lifetime, and a patient's medical history at any
284 individual institution is usually missing data from previous institutions.

285

286 Our study is also limited by the generality of some diagnosis codes and insufficient
287 disease quantification. For example, it is difficult to extract meaning from a phenotype
288 named "Abnormal findings examination of lungs" that is too general and may represent
289 undiagnosed diseases. Additionally, diagnoses do not indicate the severity of disease; a
290 diagnosis of type 2 diabetes applies to both well-managed and poorly managed disease.
291 However, in the poorly managed case, we are likely to see diagnoses representing
292 additional complications. Future studies that supplement diagnosis binary data with severity
293 data could provide insights about the relationships between disease severity and PTB. In
294 EHR, this could be addressed with lab, vitals, and/or clinical notes data.

295

296 We propose two main areas of further research resulting from this work. The first is
297 investigating lesser-known or unknown associations from our overall preterm analysis. Our
298 work suggests several hypotheses which warrant more detailed study, especially in EHR
299 systems complemented by other data sets. For instance, further work investigating the
300 gastrointestinal associations would benefit from a combined EHR and gut microbiome

301 biobank. The second area for future work is thorough investigation into the spontaneous
302 group. We found no associations in our spontaneous subgroup, suggesting that this group
303 may benefit from alternative approaches, such as dimensionality reduction and clustering,
304 which may identify subtypes in the heterogeneity. We are also eager to study events and
305 exposures during these pregnancies.

306

307 In a large EHR system, we replicated known associations between diagnosed conditions
308 and PTB, as well as finding associations of interest for further study. We found that some
309 pre-conception diagnoses were strong predictors of indicated PTB subgroup while all pre-
310 conception diagnoses were poor at predicting spontaneous PTB. We hypothesize that there
311 are strong associations between some diagnoses during pregnancy (e.g. life-threatening
312 ones such as sepsis) and spontaneous PTB. These associations would likely be particularly
313 prominent in our cohort of tertiary care individuals.

314

315 In this study, we only investigated one stratification of PTB: spontaneous/indicated. We
316 ran preliminary studies on stratifying early (<32 weeks gestational age) and late (32-36
317 weeks) PTB, but the sample size for the early preterm group (n=132) was prohibitively small
318 to produce any meaningful results on potential diagnoses associations. Future work should
319 explore associations for different stratifications such as early/late preterm, young/mid/old
320 maternal age, rural/suburban/urban maternal home, and low/middle/high maternal
321 socioeconomic status. Identifying different risk factors and pathways to PTB could lead to
322 better understanding of this heterogenous condition and to targeted and effective prevention
323 efforts.

324

325 **4 Methods**

326 **Data Selection**

327 **Birth Data**

328 We identified births using a perinatal database (PDB), which is maintained and
329 curated by obstetricians at UCSF. This database contains detailed information about each
330 delivery that takes place in the hospital and includes whether the delivery was spontaneous
331 or indicated. Newborn patient IDs in this database are linked to newborn patient IDs in the
332 EHR. The start of pregnancy was determined by subtracting gestational weeks from the
333 delivery date.

334

335 **Diagnosis Data**

336 Diagnosis information was obtained from UCSF's Observational Medical Outcomes
337 Partnership (OMOP) de-identified EHR database, using mapped newborn patient IDs from
338 the PDB. To be considered, diagnoses must have an ICD-9 or ICD-10 code and map to a
339 phecode, and have a start date prior to the start of pregnancy. Conditions were considered
340 to either be present or absent so that chronic conditions that may be recorded multiple times
341 per pregnancy would not overwhelm our results. Phecodes were truncated after the first
342 decimal point to provide an appropriate level of detail to conditions (**Figure S1**).

343 **Selection Criteria**

344 We selected our sample from all deliveries at UCSF between 2001 and 2022. To be
345 included, PDB maternal patient IDs must map to OMOP maternal patient IDs, and only one
346 record per delivery may be present (**Figure 1a**). Additionally, deliveries must be singleton,
347 have a recorded gestational age, have a recorded delivery date, and be from a birthing
348 person with at least one diagnosis prior to the start of pregnancy.

349

350 **Assigning conditions to pregnancies**

351 For our main analysis, we included multiple deliveries from the same birthing person.
352 Conditions were assigned per pregnancy in the following way: the condition must be
353 diagnosed prior to the pregnancy start but not in the six-month period following the most
354 recent delivery (**Figure 1b**). If no prior delivery was recorded, then conditions any time prior
355 to the pregnancy were included.

356

357 **Spontaneous and Indicated Preterm Definitions**

358 A team of clinicians manually marked all 10,668 pregnancies in this cohort with a PTB
359 status of “No” (this indicates a term birth), “spontaneous,” “PPROM,” “medically indicated,”
360 “Termination Iatrogenic,” or “PTL with TOCO and TERM.” Additionally, pregnancies had
361 data on gestational age, maternal age, maternal education level in years, and insurance
362 type. Using the gestational age data, 975 of the newborns were delivered at fewer than 37
363 weeks, and 9,693 newborns were delivered at 37 weeks or more (**Figure 1a**). Pregnancies
364 that were labeled with a gestational age of less than 37 weeks and a PTB status of
365 “medically indicated” or “Termination Iatrogenic” were classified as indicated for this study.
366 Pregnancies that were labeled with a gestational age of less than 37 weeks and a PTB
367 status of “spontaneous,” “PPROM,” or “PTL with TOCO and TERM” were classified as
368 spontaneous for this study.

369

370 In some cases, the PTB status value did not align with the gestational age value. In
371 these cases, the pregnancies were dropped. Pregnancies marked with a gestational age of
372 37+ weeks and a spontaneous PTB status (n=18) represent individuals who experienced
373 medically interrupted spontaneous preterm labor and delivered after term. It is unknown why
374 some pregnancies would be marked with a gestational age less than 37 weeks and “No”
375 PTB status (n=106).

376 **Diagnosis-PTB Association Analysis**

377 For all diagnoses occurring in at least one recorded birth, we implemented crude and
378 adjusted logistic regression to test the associations between each diagnosis and unstratified
379 PTB, indicated PTB, and spontaneous PTB.

380

381 **Logistic Regression**

382 Odds ratios and p-values were calculated by comparing spontaneous preterm
383 pregnancies to term controls and indicated preterm pregnancies to term controls using
384 logistic regression. We used the `glm()` function in R.

385 **Covariates**

386 Based on previous findings regarding PTB, we wanted to adjust for maternal age,
387 socioeconomic status (SES), and racism. While we have maternal age as a variable, we can
388 only use proxies for the latter two areas. For SES, we tested insurance status and maternal
389 education. For racism, we used self-reported race and ethnicity. Our final covariates were
390 maternal age, insurance (public vs private), education, and self-reported race and ethnicity.
391 A smoothing spline was applied to maternal age to capture the non-linear relationship
392 between age and PTB (29). Maternal education was reduced from the raw number of years
393 value to the categories less than 12th grade, 12th grade, or college. As missingness for
394 covariates was present in the data, unknown was considered to be a separate category for
395 each.

396 **P-Value Significance, Bootstrapping and Plotting**

397 When classifying a phenotype as significant or not significant, p-values were adjusted for
398 multiple hypothesis testing using the Benjamini Hochberg correction and tested against the
399 threshold $p_{BH} < 0.05$ for 100% of 50 bootstrap iterations. Significant phenotypes with zero and

400 infinity errors in the logistic regression (e.g. Intestinal infection, $OR=1/\infty$, $p=0$, $n_{indicated_preterm} = 0$,
401 $n_{term} = 30$) were dropped.

402

403 As most of the diagnoses occur in fewer than 3% or 250 patients, it is important to
404 consider that a small sample of individuals with a rare diagnosis could test as a significant
405 association by chance instead of a causal relationship. Bootstrapping was performed on all
406 analyses for 50 iterations of removing one instance of each diagnosis. For each iteration, a
407 unique individual was selected. If the number of instances of the diagnosis was less than 50,
408 then each instance was removed in exactly one iteration. If the number of instances of the
409 diagnosis was 50 or greater, then each instance was removed at most in one iteration.

410

411 The Manhattan and Forest plots were generated using the R packages ggplot2 (version
412 3.4.2) (30) and ggrepel (version 0.9.3) (31). The odds-odds and supplementary figures were
413 plotted using the Python packages Matplotlib (version 3.7.0) (32) and Seaborn (version
414 0.12.0) (33). Plots show adjusted p-values. Odds ratios and 95% confidence intervals are
415 not adjusted for multiple hypothesis testing.

416 **5 List of Abbreviations**

417 EHR: Electronic Health Record

418 ICD: International Classification of Diseases

419 PTB: Preterm Birth

420 SES: Socioeconomic status

421 UCSF: University of California, San Francisco

422 **6 Declarations**

423 **Ethics Approval and consent to participate**

424 This study was approved by the Institutional Review Board of University of California San
425 Francisco (#17-22929).

426

427 **Consent for Publication**

428 Not Applicable

429

430 **Availability of data and materials**

431 Overall preterm, indicated preterm, and spontaneous preterm diagnosis association
432 results are in Supplementary Files 1-3, respectively. In our association analyses, some
433 diagnoses had very low (<10) patient counts. To maintain patient de-identification, exact
434 counts, odds ratios, and p-values are redacted for those diagnoses in Supplementary Files
435 1-3. UCSF-affiliated individuals can request access to UCSF EHR data by contacting UCSF
436 Information Commons (Info.Commons@ucsf.edu).

437 The custom code/software we generated are available in the repository
438 “stratified_PTBAssociation_study” available here
439 [https://github.com/hanmochturt/stratified_PTBAssociation_study] This
440 contains instructions for OMOP EHR data queries and all of the code for patient filtering,
441 diagnosis aggregation, overall PTB association analysis, indicated and spontaneous PTB
442 association analysis, healthcare time trajectory analysis, robustness testing, and figure
443 creation.

444 **Competing interests**

445 The authors declare no competing interests.

446 **Funding**

447 We would like to acknowledge the T32 institutional training grant (5 T32DE007306) and
448 March of Dimes for funding.

449 **Authors' Contributions**

450 Conceptualization, JMC, JR, AT, TO, JAC, and MS; Methodology, JMC, HT, JR, AT, TO,
451 JAC, and MS; UCSF EHR data access, JMC; Formal Analysis, JMC and HT; Clinical
452 Insights, OY; Writing – Original Draft, JMC and HT; Writing – Review & Editing, JMC, HT,
453 JAC, and MS; Visualization, JMC and HT; all authors read and approved the final
454 manuscript.

455 **Acknowledgements**

456 We would like to acknowledge Timothy Wen for his clinical insight, Melissa Rosenstein
457 for access to the Perinatal Database, and the UCSF Information Commons team for EHR
458 data access, members of the Sirota and Capra Labs for useful discussion.

459

460 **References**

- 461 1. Walani SR. Global burden of preterm birth. Vol. 150, International Journal of
462 Gynecology and Obstetrics. 2020.
- 463 2. Behrman RE, Butler AS. Mortality and acute complications in preterm infants.
464 Preterm birth: causes, consequences & prevention. 2006;
- 465 3. Moster D, Lie RT, Markestad T. Long-Term Medical and Social Consequences of
466 Preterm Birth. New England Journal of Medicine. 2008;359(3).
- 467 4. Kramer MS, Lydon J, Goulet L, Kahn S, Dahhou M, Platt RW, et al. Maternal
468 stress/distress, hormonal pathways and spontaneous preterm birth. Paediatr Perinat
469 Epidemiol. 2013;27(3).

- 470 5. Andrews WW, Hauth JC, Goldenberg RL. Infection and preterm birth. *Am J Perinatol.*
471 2000;17(7).
- 472 6. Brou L, Almlı LM, Pearce BD, Bhat G, Drobek CO, Fortunato S, et al. Dysregulated
473 biomarkers induce distinct pathways in preterm birth. *BJOG.* 2012;119(4).
- 474 7. Jelliffe-Pawlowski LL, Baer RJ, Blumenfeld YJ, Ryckman KK, O’Brodivich HM,
475 Gould JB, et al. Maternal characteristics and mid-pregnancy serum biomarkers as
476 risk factors for subtypes of preterm birth. *BJOG.* 2015;122(11).
- 477 8. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of
478 preterm birth. Vol. 371, *The Lancet.* 2008.
- 479 9. Medically Indicated Late-Preterm and Early-Term Deliveries. *ACOG Comm Opin.*
480 2021 Jun 24;138(1).
- 481 10. Valencia CM, Mol BW, Jacobsson B, Simpson JL, Norman J, Grobman W, et al.
482 FIGO good practice recommendations on modifiable causes of iatrogenic preterm
483 birth. *International Journal of Gynecology and Obstetrics.* 2021;155(1).
- 484 11. Zhang YJ, Shen J, Lin SB, Lu C, Jiang H, Sun Y, et al. The risk factors of preterm
485 birth: A multicentre case–control survey in China in 2018. *J Paediatr Child Health.*
486 2022;58(8).
- 487 12. Purisch SE, Gyamfi-Bannerman C. Epidemiology of preterm birth. Vol. 41, *Seminars*
488 *in Perinatology.* 2017.
- 489 13. Malawa Z, Gaarde J, Spellen S. Racism as a root cause approach: A new
490 framework. *Pediatrics.* 2021;147(1).
- 491 14. Kistka ZAF, Palomar L, Lee KA, Boslaugh SE, Wangler MF, Cole FS, et al. Racial
492 disparity in the frequency of recurrence of preterm birth. *Am J Obstet Gynecol.*
493 2007;196(2).
- 494 15. Paquette AG, Hood L, Price ND, Sadovsky Y. Deep phenotyping during pregnancy
495 for predictive and preventive medicine. *Sci Transl Med.* 2020;12(527).

- 496 16. Abraham A, Le B, Kosti I, Straub P, Velez-Edwards DR, Davis LK, et al. Dense
497 phenotyping from electronic health records enables machine learning-based
498 prediction of preterm birth. *BMC Med.* 2022 Dec 1;20(1).
- 499 17. Harel Z, Park AL, McArthur E, Hladunewich M, Dirk JS, Wald R, et al. Prepregnancy
500 renal function and risk of preterm birth and related outcomes. *CMAJ.* 2020;192(30).
- 501 18. Savova GK, Masanz JJ, Ogren P V., Zheng J, Sohn S, Kipper-Schuler KC, et al.
502 Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES):
503 Architecture, component evaluation and applications. *Journal of the American
504 Medical Informatics Association.* 2010;17(5).
- 505 19. Musilova I, Pliskova L, Gerychova R, Janku P, Simetka O, Matlak P, et al. Maternal
506 white blood cell count cannot identify the presence of microbial invasion of the
507 amniotic cavity or intra-amniotic inflammation in women with preterm prelabor rupture
508 of membranes. *PLoS One.* 2017;12(12).
- 509 20. Koleva-Korkelia I. MATERNAL LEUKOCYTOSIS AS DIAGNOSTIC MARKERS IN
510 SPONTANEOUSLY DECLARED PRETERM BIRTH. *Proceedings of CBU in
511 Medicine and Pharmacy.* 2021;2.
- 512 21. Bartkeviciene D, Pilypiene I, Drasutiene G, Bausyte R, Mauricas M, Silkunas M, et al.
513 Leukocytosis as a prognostic marker in the development of fetal inflammatory
514 response syndrome. *Libyan Journal of Medicine.* 2013;8(1).
- 515 22. Bolton CE, Bush A, Hurst JR, Kotecha S, McGarvey L. Lung consequences in adults
516 born prematurely. *Vol. 70, Thorax. BMJ Publishing Group; 2015. p. 574–80.*
- 517 23. Koire A, Chu DM, Aagaard K. Family history is a predictor of current preterm birth. In:
518 *American Journal of Obstetrics and Gynecology MFM.* 2021.
- 519 24. Zhuang X, Cui AM, Wang Q, Cheng XY, Shen Y, Cai WH, et al. Liver Dysfunction
520 during Pregnancy and Its Association With Preterm Birth in China: A Prospective
521 Cohort Study. *EBioMedicine [Internet].* 2017 Dec 1;26:152–6. Available from:
522 <https://doi.org/10.1016/j.ebiom.2017.11.014>

- 523 25. Amadou C, Nabi O, Serfaty L, Lacombe K, Boursier J, Mathurin P, et al. Association
524 between birth weight, preterm birth, and nonalcoholic fatty liver disease in a
525 community-based cohort. *Hepatology* [Internet]. 2022 Nov 1;76(5):1438–51.
526 Available from: <https://doi.org/10.1002/hep.32540>
- 527 26. You S, Cui AM, Hashmi SF, Zhang X, Nadolny C, Chen Y, et al. Dysregulation of bile
528 acids increases the risk for preterm birth in pregnant women. *Nat Commun* [Internet].
529 2020;11(1):2111. Available from: <https://doi.org/10.1038/s41467-020-15923-4>
- 530 27. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic
531 health records: handling both selection bias and outcome misclassification.
532 *Biometrics*. 2022;78(1).
- 533 28. Altman MR, Colorafi K, Daratha KB. The Reliability of Electronic Health Record Data
534 Used for Obstetrical Research. *Appl Clin Inform*. 2018;9(1).
- 535 29. Hutcheon JA, Moskosky S, Ananth C V., Basso O, Briss PA, Ferré CD, et al. Good
536 practices for the design, analysis, and interpretation of observational studies on birth
537 spacing and perinatal health outcomes. *Paediatr Perinat Epidemiol*. 2019;33(1).
- 538 30. Wickham H. *ggplot2: Elegant Graphics for Data Analysis* [Internet]. Springer-Verlag
539 New York; 2016. Available from: <https://ggplot2.tidyverse.org>
- 540 31. Slowikowski K. *ggrepel: Automatically Position Non-Overlapping Text Labels with*
541 *“ggplot2”* [Internet]. 2023. Available from: [https://CRAN.R-](https://CRAN.R-project.org/package=ggrepel)
542 [project.org/package=ggrepel](https://CRAN.R-project.org/package=ggrepel)
- 543 32. Hunter JD. *Matplotlib: A 2D graphics environment*. *Comput Sci Eng*. 2007;9(3):90–5.
- 544 33. Waskom M. *seaborn: statistical data visualization*. *J Open Source Softw*. 2021 Apr
545 6;6(60):3021.
- 546

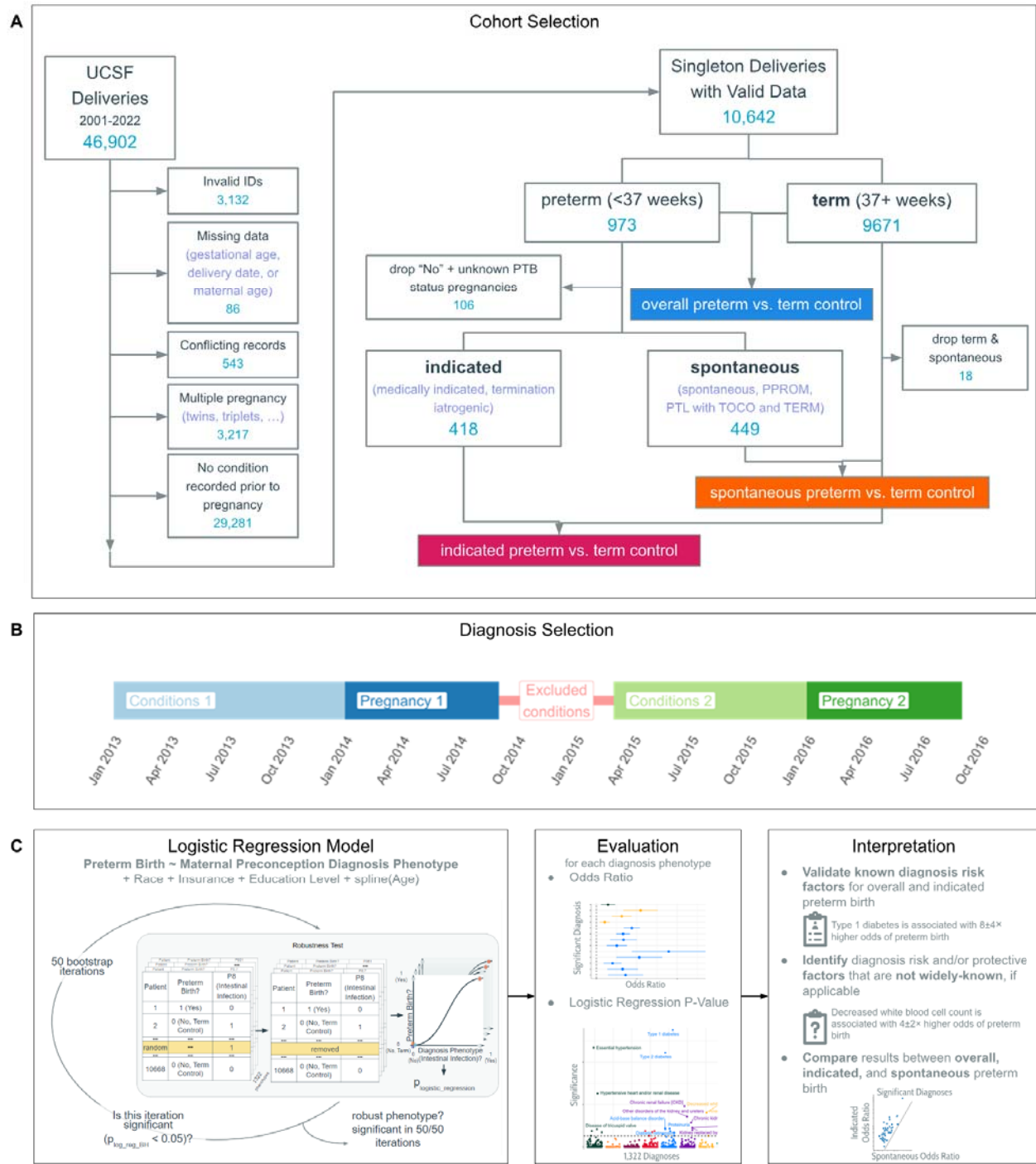


Figure 1: Schematic of the Approach for Testing Associations Between Preterm Birth and Diverse Phenotypes.

(A) Criteria for identifying the 10,642 individuals studied and assigning them to overall preterm, indicated preterm, spontaneous preterm, and term groups for subsequent logistic regression analyses. (B) Diagnoses before conception are used in this study. For a person's first birth (or only birth), diagnoses are recorded from the start of their record until the start of conception. If multiple births are recorded for the same individual, diagnoses for subsequent births are recorded starting 6 months from after the previous delivery to the start of the next conception.

(C) Overview of the logistic regression analysis, covariates, evaluation, and interpretation for associations between preterm birth and the 1322 diagnosis phenotypes considered. P-values were adjusted for multiple hypothesis testing and a permutation test was used to ensure associations were robust.

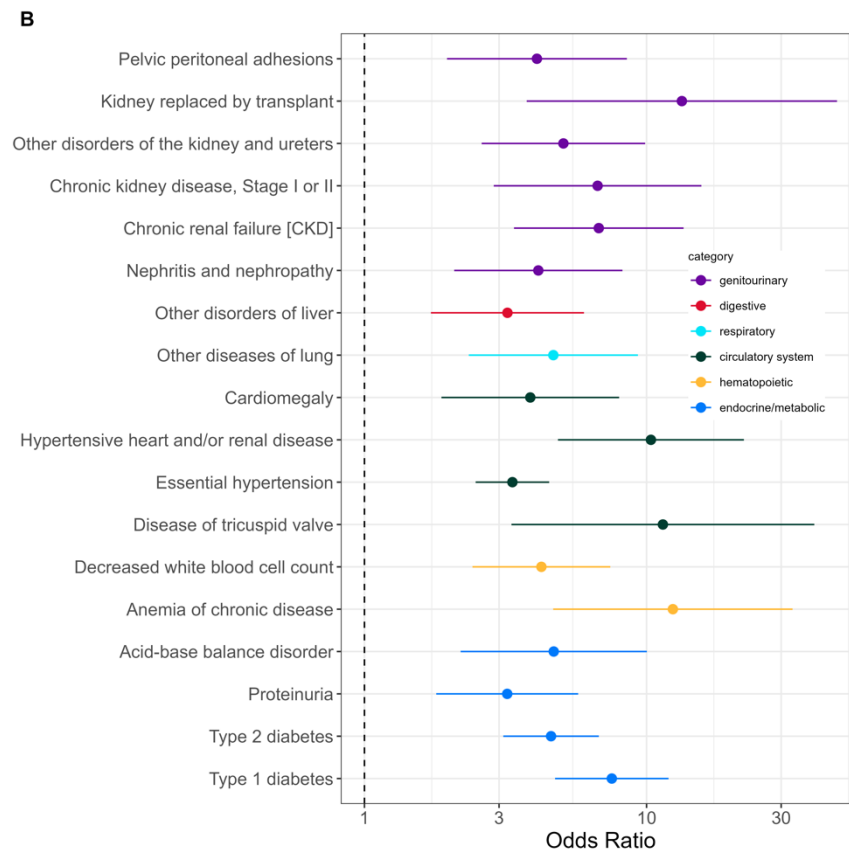
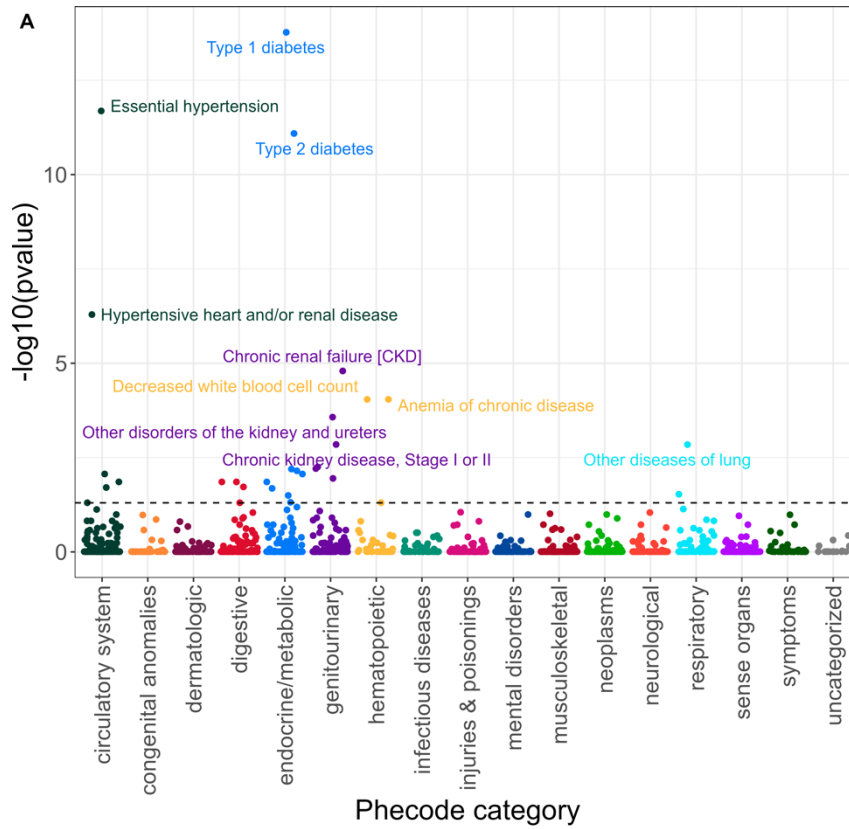


Figure 2: Testing Associations Between Clinical Phenotypes and PTB Identifies Known Risk Factors and Novel Candidates.

(A) P-values from logistic regression tests of the association of 1322 clinical phenotypes with preterm (n=973) vs. term births (n=9671). Seventeen phenotypes passed the Benjamini Hochberg multiple testing corrected false discovery rate threshold of 5% (dashed line) and were robust to small changes in the data set. Phenotypes were represented as phecodes and plotted by phecode category. Significant, robust associations are labelled. **(B)** Forest plot shows odds ratios and 95% confidence intervals of the 17 conditions that were significantly and robustly associated with overall PTB. “Other disorders of liver” represents conditions including liver lesion, liver cirrhosis, liver mass, liver carcinoma, and fatty liver. “Other diseases of lung” represents conditions including lung consolidation, interstitial lung diseases, and lung mass.

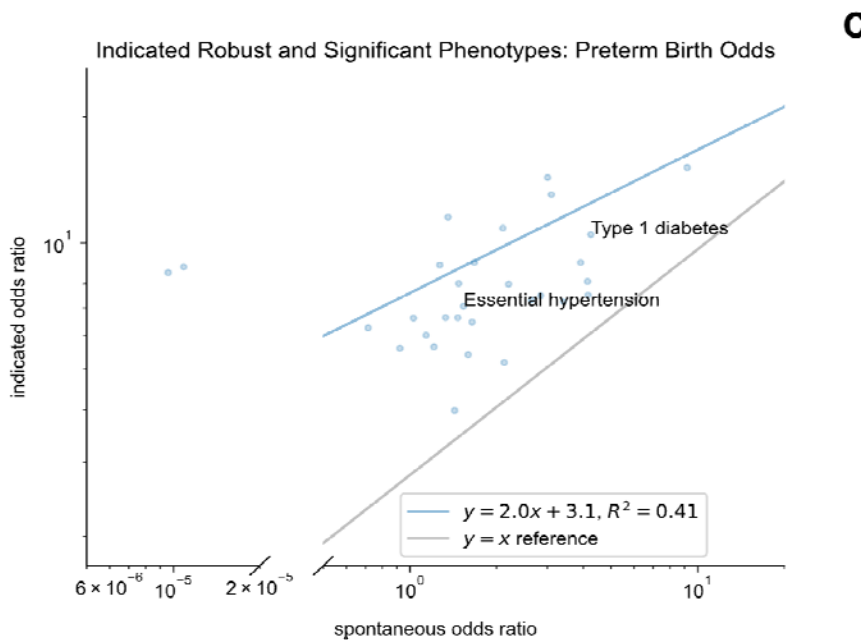
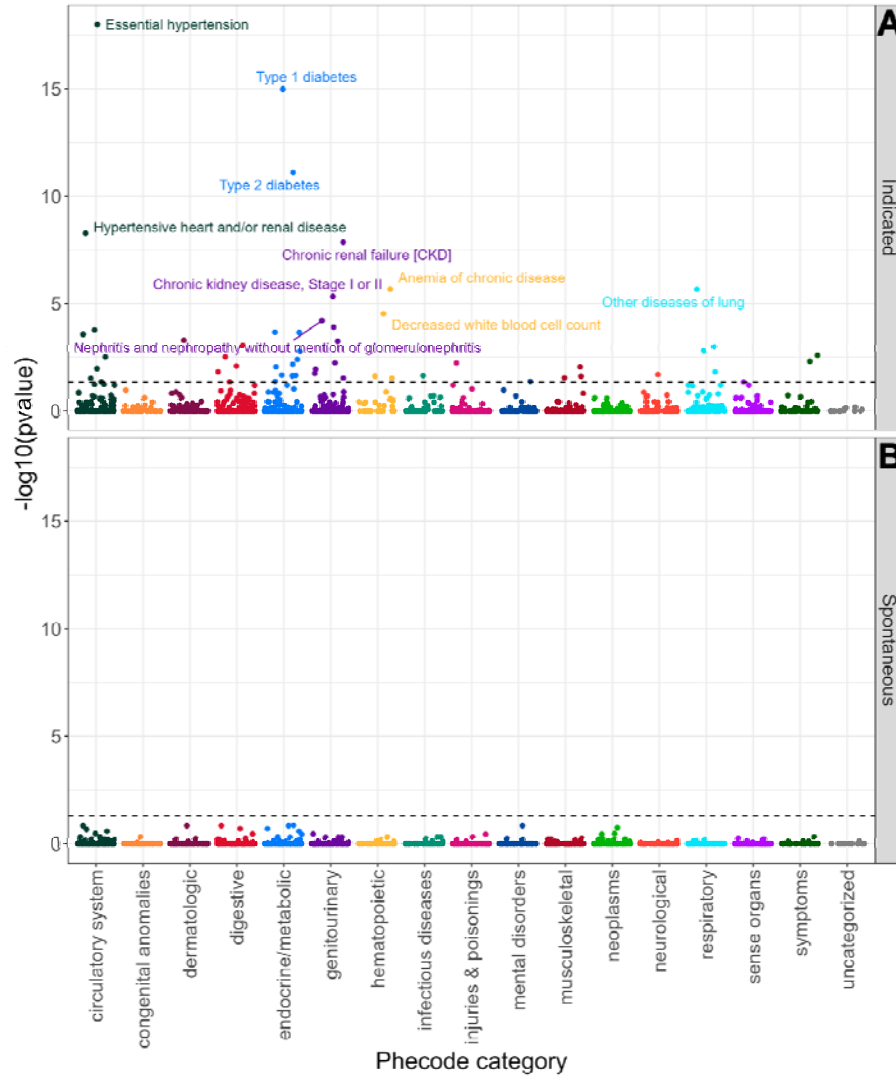


Figure 3: Many conditions associate with risk for indicated preterm birth, but none with spontaneous preterm birth.

(A) P-values from logistic regression tests of the association of 1322 clinical phenotypes with indicated preterm (n=418) vs. term births (n=9671). Thirty phenotypes passed the multiple testing corrected Benjamini Hochberg threshold of 5% (dashed line) and were robust to small changes in the data set. (B) P-values from logistic regression tests of the association of 1322 clinical phenotypes with spontaneous preterm (n=449) vs. term births (n=9671). No phenotypes significantly associated with spontaneous preterm birth. (C) Comparison of the odds ratios for the 30 phenotypes significantly associated with indicated preterm birth between tests for indicated and spontaneous preterm birth. The odds ratios are correlated ($r^2=0.42$, linear regression, left outliers dropped), but the relationships have systematically lower magnitude in the spontaneous cohort. The two most significant indicated phenotypes are labeled.

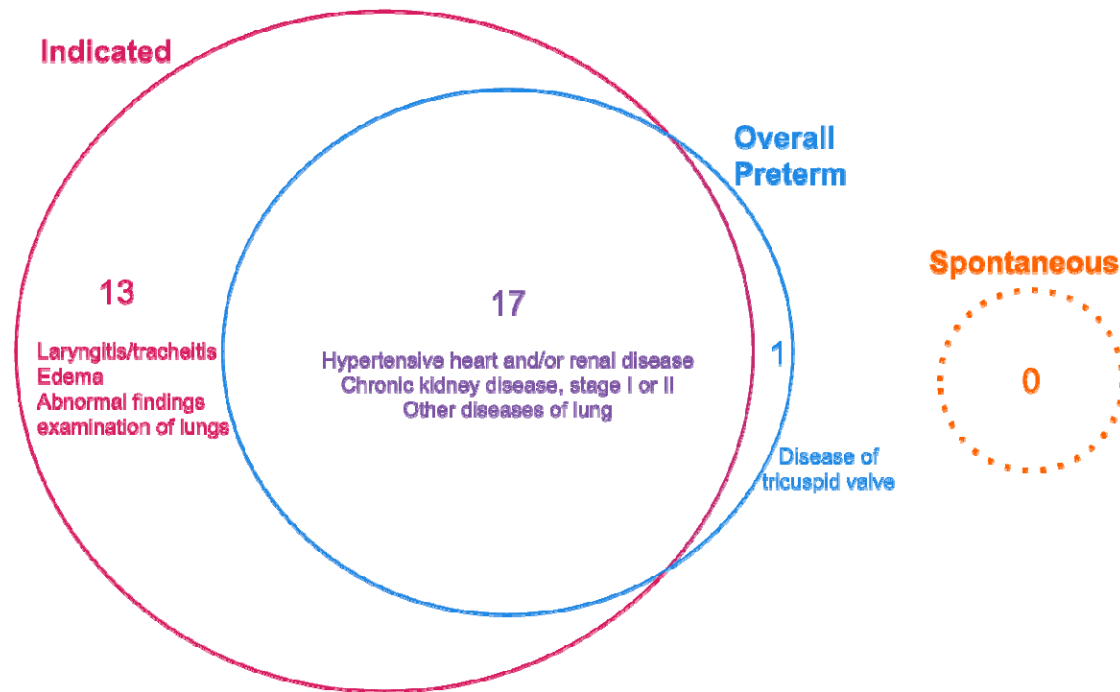


Figure 4: The strongest associations with overall preterm birth are also associated with indicated preterm birth, but not spontaneous preterm birth.

Many kidney, cardiac, liver, and pulmonary conditions and diabetes are associated with overall and/or indicated preterm birth. Disease of tricuspid valve is a strong risk factor in overall preterm birth but not as strong in indicated preterm birth.

	Indicated (N=418)	Spontaneous (N=449)	All Preterm (N=973)	Term (N=9671)	Overall (N=10643)
Race					
1:Single race-White	171 (40.9%)	205 (45.7%)	426 (43.8%)	4715 (48.8%)	5141 (48.3%)
2:Single race-Black	55 (13.2%)	44 (9.8%)	108 (11.1%)	523 (5.4%)	631 (5.9%)
3:Single race-Latina	57 (13.6%)	46 (10.2%)	111 (11.4%)	679 (7.0%)	790 (7.4%)
4:Single race-Asian/Pacific Islander	79 (18.9%)	100 (22.3%)	200 (20.6%)	2466 (25.5%)	2666 (25.0%)
5:Single race-Native		1 (0.2%)	1 (0.1%)	13 (0.1%)	14 (0.1%)
6:Mul race-Latina + other race	23 (5.5%)	17 (3.8%)	48 (4.9%)	433 (4.5%)	481 (4.5%)
7:Mul race-other races	10 (2.4%)	11 (2.4%)	23 (2.4%)	253 (2.6%)	276 (2.6%)
8:Other race	17 (4.1%)	16 (3.6%)	36 (3.7%)	426 (4.4%)	462 (4.3%)
9:Unknown	6 (1.4%)	9 (2.0%)	20 (2.1%)	162 (1.7%)	182 (1.7%)
Maternal age					
Mean (SD)	33.9 (6.1)	34.0 (5.4)	34.0 (5.64)	34.5 (4.84)	34.4 (4.92)
Median [Min, Max]	35.0 [15.0,51.0]	35.0 [14.0,55.0]	35.0 [15.0, 54.0]	35.0 [14.0, 55.0]	35.0 [14.0, 55.0]
Missing	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)	1 (0.0%)
Private Insurance					
0:No	54 (12.9%)	40 (8.9%)	102 (10.5%)	579 (6.0%)	681 (6.4%)
1:Yes	363 (86.8%)	407 (90.6%)	865 (88.9%)	9050 (93.6%)	9915 (93.2%)
unknown	1 (0.2%)	2 (0.4%)	6 (0.6%)	41 (0.4%)	47 (0.4%)
Maternal Education					

<12 years	13 (3.1%)	11 (2.4%)	25 (2.6%)	71 (0.7%)	96 (0.9%)
12 years	124 (29.7%)	76 (16.9%)	218 (22.4%)	1348 (13.9%)	1566 (14.7%)
>12 years	230 (55.0%)	308 (68.6%)	589 (60.5%)	7550 (78.1%)	8139 (76.5%)
unknown	51 (12.2%)	54 (12.0%)	141 (14.5%)	702 (7.3%)	843 (7.9%)
Diagnosis Count					
Number of unique diagnoses Mean (SD)	10.9 (13.4)	8.8 (10.1)	9.6 (11.8)	7.3 (8.5)	7.5 (8.9)
Median [Min, Max]	6.0 [1.0,101.0]	5.0 [1.0,88.0]	5.0 [1.0,101.0]	5.0 [1.0,118.0]	5.0 [1.0,118.0]
Diagnosis Time					
Medical diagnosis visit time before conception Mean (SD)	1.8 (1.6)	1.7 (1.6)	1.7 (1.6)	1.5 (1.6)	1.6 (1.6)
Median [Min, Max]	1.3 [0.0,8.7]	1.1 [0.0,9.0]	1.2 [0.0,9.0]	1.0 [0.0,21.7]	1.0 [0.0,21.7]
Medical diagnosis first visit time before conception Mean (SD)	2.4 (1.9)	2.1 (1.9)	2.3 (1.9)	1.9 (1.8)	2.0 (1.9)
Median [Min, Max]	2.0 [0.0,8.7]	1.5 [0.0,9.0]	1.7 [0.0,9.0]	1.3 [0.0,21.7]	1.4 [0.0,21.7]

Table 1: Demographics

Race, maternal age, maternal education level, insurance status (private, public, or unknown), and diagnosis distributions of individuals included in this study.

SUPPLEMENTAL FIGURES

7 Supplementary Information

1.1 Supplemental Figures

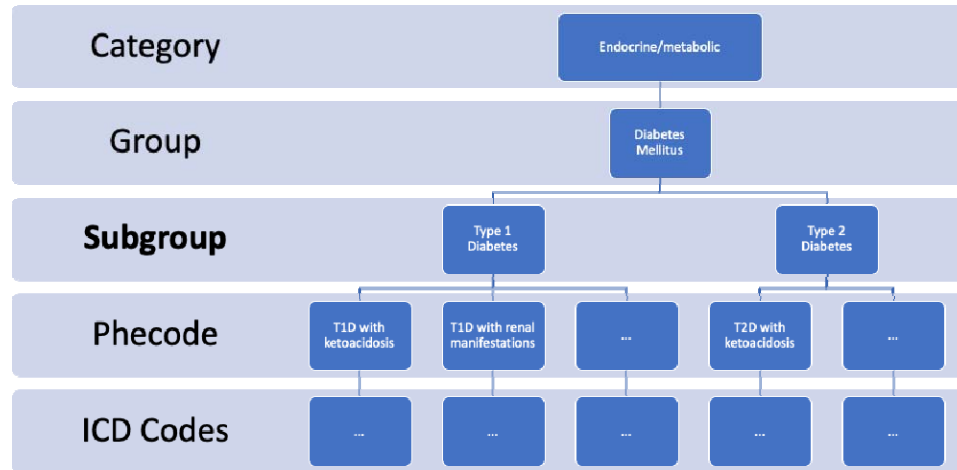


Figure S1: Phecode subgroups used to define diagnosis phenotypes

We use phecode subgroups to define diagnosis phenotypes. This allows us to test diagnosis-preterm associations with sufficient detail.

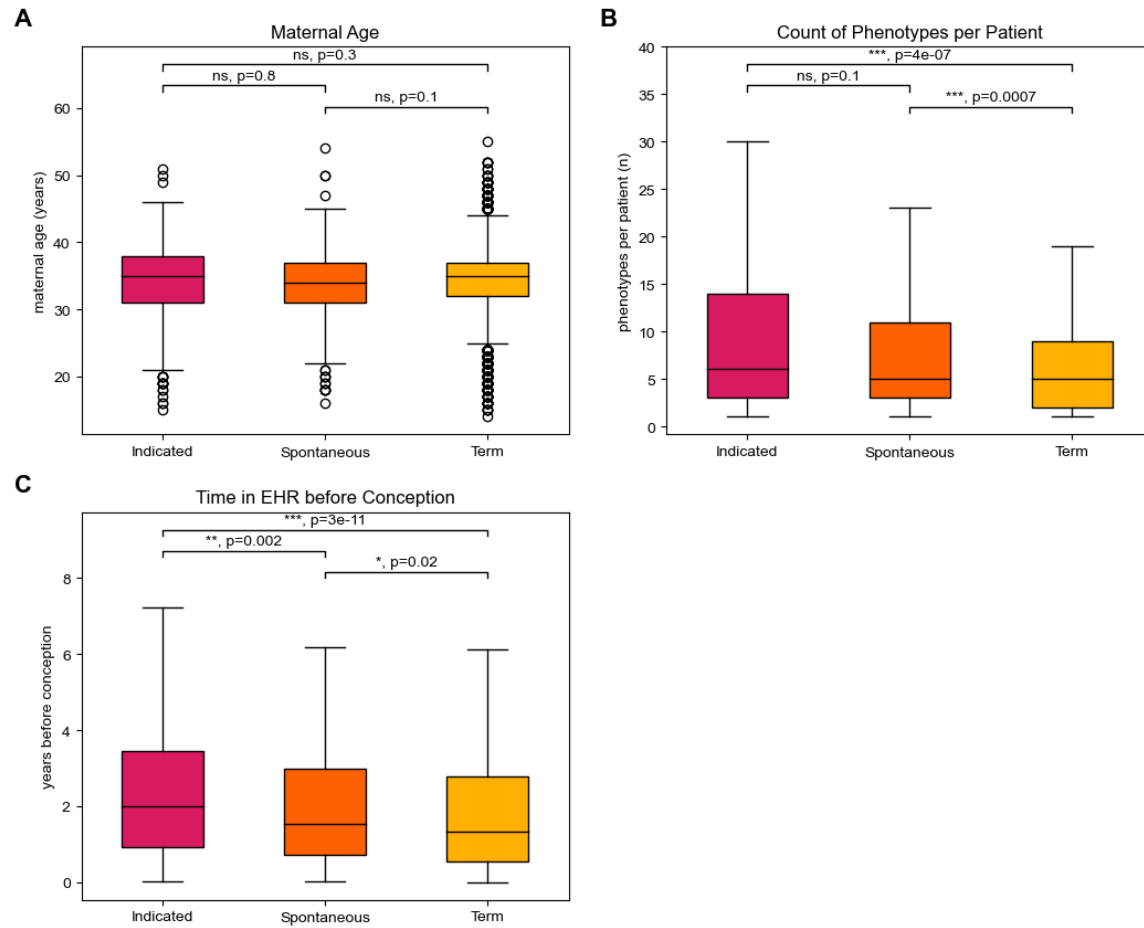


Figure S2: Patient Distributions for Maternal Age, Number of Diagnoses per Patient, First Visit per Patient, All Visits per Patient

(Ai) By 2-sided Mann-Whitney U rank test, there are no significant differences in maternal age between indicated, spontaneous, and term. **(B)** Additionally, by the same statistics test, outliers dropped, the spontaneous preterm and indicated preterm patients have significantly more diagnoses than the term patients. **(C)** Indicated individuals have the longest (time) EHR length, followed by spontaneous individuals, then term individuals with the lowest.

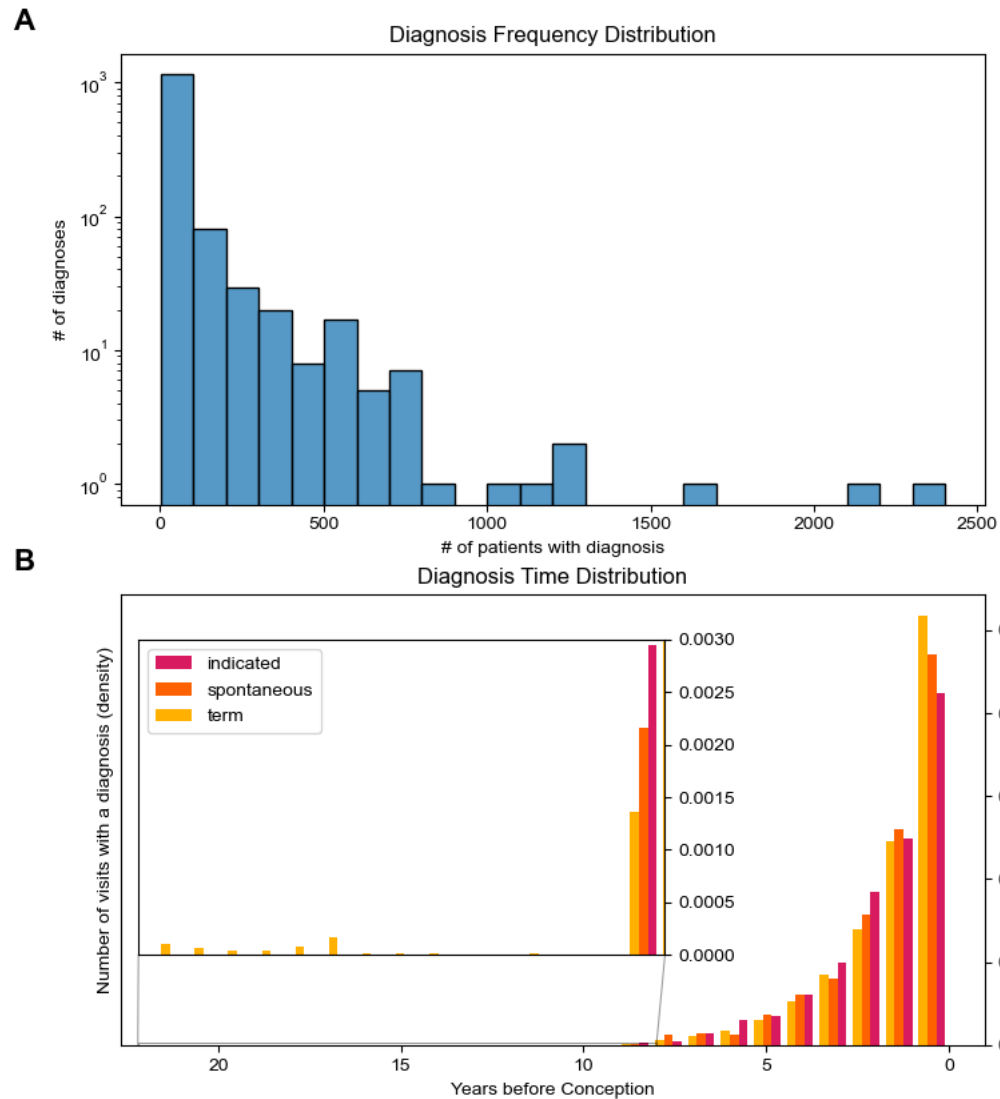


Figure S3: Diagnoses Frequency Distribution

(A) Most (1,148 of 1,322) diagnosis phenotypes occur rarely (in fewer than 100/10642 patients). **(B)** Most medical visits with a diagnosis occurred within 2 years before conception.