HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2023 January 01.

Published in final edited form as:

Pac Symp Biocomput. 2023 ; 28: 209–220.

Using Association Rules to Understand the Risk of Adverse Pregnancy Outcomes in a Diverse Population

Hoyin Chu[1], **Rashika Ramola**[1], **Shantanu Jain**[1], **David M. Haas**[2], **Sriraam Natarajan**[3], **Predrag Radivojac**[1]

[1]Northeastern University,

[2]Indiana University School of Medicine,

[3]University of Texas at Dallas

Abstract

Racial and ethnic disparities in adverse pregnancy outcomes (APOs) have been well-documented in the United States, but the extent to which the disparities are present in high-risk subgroups have not been studied. To address this problem, we first applied association rule mining to the clinical data derived from the prospective nuMoM2b study cohort to identify subgroups at increased risk of developing four APOs (gestational diabetes, hypertension acquired during pregnancy, preeclampsia, and preterm birth). We then quantified racial/ethnic disparities within the cohort as well as within high-risk subgroups to assess potential effects of risk-reduction strategies. We identify significant differences in distributions of major risk factors across racial/ethnic groups and find surprising heterogeneity in APO prevalence across these populations, both in the cohort and in its high-risk subgroups. Our results suggest that risk-reducing strategies that simultaneously reduce disparities may require targeting of high-risk subgroups with considerations for the population context.

Keywords

Adverse pregnancy outcomes; risk assessment; health disparities

1. Introduction

The U.S. department of Health and Human Services defines health disparity as a particular kind of health difference that is closely linked with social, economic, and/or environmental disadvantage.[1] The American healthcare system has many examples of disparities between communities.[2–4] In 2016–2018, the all-cause mortality rate among Black populations was 24% higher than among White populations nationally.[5] Similarly, the Hispanic population in the USA has lesser access to health insurance than other racial/ethnic groups—before the implementation of the Affordable Care Act in 2014, 30% of Hispanic individuals reported no health insurance as compared to 11% of non-Hispanic White individuals.

In addition to the adverse consequences for the affected people and their communities, health disparities result in larger economic burden for the entire nation.[6,7] Eliminating health disparities could have reduced direct medical expenses by approximately $230 billion, and

indirect productivity costs by more than $1 trillion for the years 2003–2006, with the most of the estimated cost reduction attributed to the generally poorer health outcomes of the Black and Hispanic communities.[6]

Adverse pregnancy outcomes (APOs) such as gestational diabetes mellitus (GDM), preeclampsia (PReEc), preterm birth (PTB) and new hypertension (NewHTN) are known to disproportionally affect racial/ethnic groups. As an example, a study of 5,562 women found the rate of GDM was the highest among Asian American women (16%), followed by non-Hispanic Black women (9%), Hispanic women (11%), and non-Hispanic White women (8%).[8] In another study, non-Hispanic Black women were found to be significantly more likely to experience preterm birth, hypertensive disease of pregnancy, and small-for-gestational-age birth than were non-Hispanic White women.[9] Understanding these disparities is critical to ensuring equitable health outcomes; however, due to the complex interaction between biological, social, and environmental factors, the mechanisms that lead to their formation are difficult to identify. It therefore remains challenging to design policies or intervention strategies that can reduce both APO risks and existing disparities.[10]

When designing this study, we had four different goals in mind. First, to identify subgroups at high risk for APOs from a large cohort pregnant women. Second, to quantitatively measure racial/ethnic disparities within these high-risk subgroups and compare them to the population-level disparities. Third, to identify potential intervention strategies that may lead to the greatest reduction in APO prevalence. And fourth, to measure the impact of such intervention strategies on existing disparities. To achieve this, we obtained data from the diverse nuMoM2b cohort which contained clinical data for 10,038 nulliparous women,[11] and used association rule mining to identify high-risk subgroups. By increasing the resolution of the disparity analysis from the population-level to high-risk subgroups, we gained additional insights into the interplay between the main risk factors and disproportionate health outcomes. In addition, by measuring the effects of potential intervention on disparity, we found that the largest risk-reducing intervention may not be the largest disparity-reducing intervention. This finding could have implications for the design of future clinical interventions, as risk factors may vary significantly across racial/ethnic groups.

2. Methods

2.1. The nuMoM2b cohort

The Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-To-Be (nuMoM2b) cohort was recruited prospectively to identify factors that contribute to APOs.[11] The study enrolled 10,038 subjects from eight clinical centers in the US. Women were eligible for enrollment if they had a viable singleton gestation, had no previous pregnancy that lasted more than 20 weeks of gestation (i.e., nulliparous), and were between 6 0/7 and 13 6/7 weeks of gestation at enrollment, which was also the first study visit. Haas et al.[11] provide an overview of the biospecimen collection, clinical measurements, and standardized questionnaire instruments that were collected during each of the three study visits and at delivery. The cohort is racially and ethnically diverse, with more than 4,000 individuals reporting race other than White, and has a high concordance between self-reported race and

inferred ancestry from genetic data.[12] Operationally, the cohort comprises of 1,509 subjects positive for at least one APO. Of those, 807 were positive for PTB, 568 for preeclampsia, 55 for fetal demise, 414 for GDM, and 406 experienced fetal growth restriction.

To capture an accurate representation of the participants prior to any clinical interventions, we used data from the first study visit only. For quality control, 10 individuals with high information missingness were excluded. To ensure our findings are based on sufficiently large sample sizes and to reduce possible confounding introduced by mixed cultural effects within groups, only self-reported races/ethnicities with more than 100 participants were included and participants who did not report race/ethnicity ($n = 639$) or reported more than one race ($n = 486$) were excluded. Participants were then assigned to one of four racial/ethnic groups based on their self-reported race and ethnicity: non-Hispanic Asian, non-Hispanic Black, non-Hispanic White, and Hispanic. In total, 8,903 participants were included in the final analysis.

Our study primarily relies on clinical variables and the features selected for analysis include basic demographic features and a curated set of features previously known to affect the likelihood of developing APOs.[13] These include age, body mass index (BMI), family history of diabetes mellitus (Family DM), polycystic ovary syndrome history (PCOS), Alternate Healthy Eating Index-2010 (AHEI2010) score, activity levels measured by the metabolic equivalent of tasks (METs),[14] and high blood pressure (High BP). The diet of a participant was considered “poor” if her AHEI2010 score was below the 25th percentile of all scores, “normal” if it was between 25th and 75th percentile, and “good” if it was above the 75th percentile. Consistent with previous studies, a participant’s exercise level was considered “inactive” if her METs is below 450 and “active” otherwise.[14,15] Participants reporting age or BMI of zero were recoded as having missing age or BMI. For compatibility with downstream association rules analysis, age and BMI were discretized into intervals as defined by the nuMoM2b study.[11]

2.2. Clinical data as a transactional database

To find interesting and interpretable patterns in the nuMoM2b data, we converted it to a transactional database and performed association rule mining.[16] An association rule is a probabilistic implication discovered from a transactional database. For example, in the context of nuMoM2b, a high-confidence rule {Race = Asian, Age > 40} ⇒ {GDM = 1} has the interpretation “Pregnant Asian women above the age of 40 are likely to be diagnosed with GDM”.

A transactional database $\boldsymbol{D} = \{t_1, t_2, \ldots, t_m\}$ is a set of transactions, where each transaction is a subset of items from $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$. To represent a clinical database as a transactional dataset, we first convert the collected descriptors and clinical measurements into clinically relevant binary features such as {Race = Asian}, {Age > 40} and {GDM = 1}. Then for each subject in the cohort, we create a transaction containing only those binary features (as items) that are true for the subject. For example, based on the three features above, an Asian participant above the age of 40 and diagnosed with GDM, would be represented as {Race = Asian, Age > 40, GDM = 1}. In total, 25 binary features were created from Age (5), BMI

(5), Family DM (2), PCOS (2), High BP (2), Exercise (2), Diet (3) and APOs (4) in the nuMoM2b data.

2.3. Association rules

For a transactional database $\boldsymbol{D}$ defined on a set of items $\mathcal{I}$, an association rule is an implication of the form $A \Rightarrow B$, where A and B are disjoint subsets of $\mathcal{I}$ and are referred to as the antecedent and the consequent of the rule, respectively. Typically, the evidence of a rule in $\boldsymbol{D}$ is quantified in terms of the *confidence* defined as fraction of transactions containing all items in B out of the transactions that contain all items in A. In other words, it quantifies the conditional probability of seeing B in a transaction given that A has been already seen. Formally,

$$\text{Confidence}_{\boldsymbol{D}}(A \Rightarrow B) = \frac{\text{Support}_{\boldsymbol{D}}(A \cup B)}{\text{Support}_{\boldsymbol{D}}(A)},$$

where $\text{Support}_{\boldsymbol{D}}(A)$ is the fraction of transactions in $\boldsymbol{D}$ that contain A; i.e., $\text{Support}_{\boldsymbol{D}}(A) = |\boldsymbol{D}_A|/|\boldsymbol{D}|$, and $\boldsymbol{D}_A = \{t | A \subseteq t, t \in \boldsymbol{D}\}$ is the set of transactions containing A. Applying these definitions to the example above, $\text{Confidence}_{\boldsymbol{D}}(\{\text{Race} = \text{Asian}, \text{Age} > 40\} \Rightarrow \{\text{GDM} = 1\})$ is the fraction of women diagnosed with GDM out of all Asian women above the age of 40 in the cohort; i.e., the empirical probability that a pregnant Asian woman above the age of 40 has GDM. $\text{Support}_{\boldsymbol{D}}(\{\text{Race} = \text{Asian}, \text{Age} > 40, \text{GDM} = 1\})$ is the fraction of Asian women above the age of 40 diagnosed with GDM in our cohort.

Association rules can be efficiently discovered with the Apriori algorithm.[16] We apply Apriori to the transactional database created from the nuMoM2b data using the efficient-apriori Python package with the parameters min support = 0.0005, min confidence = 0.001, and max length = 6. Afterwards, we extracted rules with APOs as the consequent; i.e., {GDM = 1}, {NewHTN = 1}, {PReEc = 1}, and {PTB = 1}.

2.3.1. Measuring clinical significance of association rules—

While confidence is easily interpretable as a conditional probability, it fails to capture the relative improvement over the baseline probability of the consequent.[17] Any rule $A \Rightarrow B$, where B has low support, is likely to have low confidence, irrespective of the relative increase in the conditional probability over the baseline. Such rules are still important in clinical applications; e.g., finding causal attributes for rare diseases. To overcome the limitations of confidence, we use positive likelihood ratios (LR^+), a standard measure used in clinical settings.[18] Formally,

$$\text{LR}^+(A \Rightarrow B) = \frac{\text{Confidence}_{\boldsymbol{D}}(A \Rightarrow B)/(1 - \text{Confidence}_{\boldsymbol{D}}(A \Rightarrow B))}{\text{Support}_{\boldsymbol{D}}(B)/(1 - \text{Support}_{\boldsymbol{D}}(B))},$$

with asymmetric 95% confidence intervals determined by bootstrapping.[19] We additionally test the null hypothesis that the association between A and B occurs by chance, using Fisher's exact test, and compute the p-value.

2.4. Quantitative measure of disparity

Disparity of outcomes across different groups can be measured in several ways and there is not a single best quantitative measure for it.[20] We adopt the measure often used in the field of economics to study income inequalities, and define disparity as the Gini coefficient of APO prevalence rates among different populations.[21] More formally, let a binary outcome variable Y (e.g., GDM) take values $\mathcal{Y} = \{0, 1\}$, where 1 (0) indicates presence (absence) of an APO. Let X be a variable of interest (e.g., racial/ethnic group) taking values in $\mathcal{X}$, where different values of X characterize different subpopulations of interest. Let $p(x, y)$ be a joint distribution over variables X and Y. We define the disparity of Y with respect to (w.r.t.) X as the Gini coefficient of the conditional probabilities $p(Y = 1|X = x)$ over all values of $x \in \mathcal{X}$; i.e.,

$$\delta(Y \mid X) = \text{Gini}(\{p(Y = 1 \mid X = x)\}_{x \in \mathcal{X}}), \text{ where } \text{Gini}(S) = \frac{\sum_{a \in S} \sum_{b \in S} |a - b|}{2|S| \sum_{a \in S} a}$$

computes the Gini coefficient of the set S. Note that Gini coefficient is scale independent, due to normalization by $\sum_{a \in S} a$, unlike measures such as standard deviation. This property makes it ideal to compare disparity between two populations (e.g., before and after removing high-risk individuals) with outcomes on different scales.

We study disparity of APOs w.r.t. racial/ethnic groups in the nuMoM2b dataset. Under the disparity formulation given above, an APO ∈ {GDM, PReEc, PTB, NewHTN} serves as Y and racial/ethnic groups serve as X. Let $\boldsymbol{D}$ denote a cohort under study given as a transactional database defined on a set of items $\mathcal{I}$ (Section 2.2). In particular, $\mathcal{I}$ contains items $Y = 1$ and $X = x$ for $\forall x \in \mathcal{X}$. $\boldsymbol{D}$ defines an empirical distribution over X and Y given by

$$p_{\boldsymbol{D}}(x, y) = \begin{cases} \text{Support}_{\boldsymbol{D}}(X = x \cup Y = 1) & \text{when } y = 1, \\ \text{Support}_{\boldsymbol{D}}(X = x) - \text{Support}_{\boldsymbol{D}}(X = x \cup Y = 1) & \text{when } y = 0, \end{cases}$$

where $\text{Support}_{\boldsymbol{D}}(A)$ denotes the Support of an itemset $A \subseteq \mathcal{I}$ computed on $\boldsymbol{D}$. Furthermore, the conditional probability $p_{\boldsymbol{D}}(Y = 1|X = x)$ under $\boldsymbol{D}$ is given by

$$p_{\boldsymbol{D}}(Y = 1 \mid X = x) = \text{Confidence}_{\boldsymbol{D}}(X = x \Rightarrow Y = 1),$$

where $\text{Confidence}_{\boldsymbol{D}}(A \Rightarrow B)$ denotes the confidence of the rule $A \Rightarrow B$ computed on $\boldsymbol{D}$. Thus the disparity of the APOs (Y) w.r.t. racial/ethnic groups (X) on $\boldsymbol{D}$ is given by

$$\delta_{\boldsymbol{D}}(Y \mid X) = \sigma(\{p_{\boldsymbol{D}}(Y = 1 \mid X = x)\}_{x \in \mathcal{X}}), \text{ where } \sigma(S) = \text{Gini}(S).$$

We are interested in the contribution of the high-risk subgroups, defined in terms of the risk factors such as age and BMI, towards the overall prevalence and the disparities of each APO. To do so, we evaluate the relative difference in APO prevalence and disparity when the high-risk participants are omitted from the cohort. Let $R \subseteq \mathcal{I}$ be the attributes (not including

APO or racial/ethnic groups) identifying the high-risk individuals. Let $\boldsymbol{D}_R = \{t | R \subseteq t, t \in \boldsymbol{D}\}$ denote the set of transactions (individuals) in $\boldsymbol{D}$ that contain R. Let $\overline{\boldsymbol{D}}_R = \boldsymbol{D} \backslash \boldsymbol{D}_R$ be the set of transactions in $\boldsymbol{D}$ that do not contain R. The disparity of the APOs (Y) w.r.t. racial/ethnic groups (X) on $\overline{\boldsymbol{D}}_R$ is given by,

$$\delta_{\overline{\boldsymbol{D}}_R}(Y \mid X) = \sigma\left(\left\{p_{\overline{\boldsymbol{D}}_R}(Y = 1 \mid X = x)\right\}_{x \in \mathcal{X}}\right).$$

The relative change in disparity on removing the participants having all phenotypes/attributes in R is given by

$$\frac{\delta_{\overline{\boldsymbol{D}}_R}(Y \mid X) - \delta_{\boldsymbol{D}}(Y \mid X)}{\delta_{\boldsymbol{D}}(Y \mid X)}.$$

Similarly, for the subpopulation having $X = x$, the relative change in the APO prevalence rate on removing the participants having all phenotypes/attributes in R is given by

$$\frac{p_{\overline{\boldsymbol{D}}_R}(Y = 1 \mid X = x) - p_{\boldsymbol{D}}(Y = 1 \mid X = x)}{p_{\boldsymbol{D}}(Y = 1 \mid X = x)}.$$

2.4.1. Identifying high-risk subgroups—To identify high-risk subgroups used in the disparity analysis, we started with the initial set of rules with the APOs in the consequent, that pass the support and confidence thresholds. The rules were further filtered based on the following inclusion criteria: LR^+ value above 1; does not contain the variable of interest (race/ethnicity) in the antecedent; and the size of the antecedent is no more than 3.

3. Results

3.1. Association rules effectively identify high-risk subgroups

A total of 1,627 rules satisfied filtering criteria, among which 726 were nominally significant ($p < 0.05$) and 527 (GDM: 188; NewHTN: 130; PReEc: 119; PTB: 90) were significant after adjusting for multiple hypothesis testing using the Benjamini-Hochberg procedure.[22] Among the statistically significant subgroups, 21 rules had one attribute in the antecedent, 146 rules had two attributes and 360 had three attributes. BMI and Age were the two most common attributes in the rules, where 339 rules (64.3%) contained a BMI attribute and 234 rules (44.4%) contained an Age attribute (Table S1).

The generated rules were able to capture many known risk factors that are common to all APOs. For example, obesity is a known risk factor for APOs and the subgroup {BMI ≥ 35} was generated as a high-risk subgroup with varying likelihood ratios in APOs (Table 1). In addition, the generated rules were also able to capture APO-specific high-risk subgroups. For example, older age is a risk factor for GDM[23] and NewHTN,[24] while younger age is a risk factor for PTB[25] and PReEc.[26] Consistently with prior findings, we observe the corresponding risk groups {Age = 35–39} and {Age < 18} being generated in the association rules. The association between dietary choices and risk on PReEc was recently

reported[27] and we similarly see an increased risk for PReEc for the subgroup that has poor diet.

Furthermore, association rules were able to identify high-risk subgroups from combinations of features where each feature individually may not necessarily be a strong risk factor. Such combinations of features also allow for investigating the impact of a singular feature on an existing subgroup. All generated rules are listed in Supplementary Table S1, which is available online at the project github (https://github.com/hoyinchu/PSB_2023_Supplement).

3.2. Disparity is highly heterogeneous within and across APOs

We assessed the level of disparity over the entire cohort as well as in high-risk subgroups finding significant heterogeneity across APOs (Fig. 1, Table 2) and risk groups (Table S1). For example, Black participants have the lowest prevalence of GDM compared to other groups (3.1%), but the highest rates of all other APOs (9.3% in PReEc, 11.1% in PTB, 19.4% in NewHTN). Asian participants have the highest rate of GDM (10.8%), while also having the lowest rate of PReEc (3.2%). The rates of APOs in White participants are comparable to those in Hispanic, except for NewHTN (17.2% vs. 10.7%). Surprisingly, disparities in high-risk subgroups do not follow a regular pattern either. In GDM, for example, the disparity of the {Age = 35–39} subgroup ($LR^+ = 2.5$; $p = 4.7 \times 10^{-7}$) is reduced from 0.268 (population; Table 2) to 0.112 (high-risk subgroup; Table S1), whereas the disparity of the {BMI = 30–35} subgroup ($LR^+ = 1.9$; $p = 1.1 \times 10^{-6}$) is increased to 0.356 (Table S1). Similar patterns were observed in other APOs.

3.2.1. Disparities in high-risk GDM subgroups—For simplicity and interpretability, we focus our analysis mainly on single-attribute high-risk subgroups. In GDM, the {Age ≥ 40} subgroup has the highest LR^+ compared to other single-attribute subgroups, followed by the {Age = 35–39}, and {High BP = 1} subgroups; Fig. 2a. Among these subgroups, the one with the highest disparity measure was also the {Age ≥ 40} subgroup, followed by the {BMI = 30–35} and {High BP = 1} subgroups (Table S3). We then evaluated the proportion of GDM patients in each of these subgroups to understand how these risk-factors may differentially impact races/ethnicities. We found that across risk-factors, Asian participants have higher rates of GDM compared to other races/ethnicities within the same subgroup except in the {Age = 35–39} subgroup (Table S4). In particular, the rate of GDM is considerably higher in the {Age ≥ 40} subgroup, which is also the subgroup with the highest GDM disparity measure (Table S3).

We next investigated the contribution of GDM rates from each high-risk subgroup to the overall GDM rate in the cohort by calculating the relative difference between the rate of GDM before and after the subgroup is removed from the cohort; see Methods. We observe the largest decrease in GDM rate if the {Family DM = 1} subgroup is omitted, followed by {BMI ≥ 35} and {BMI = 30–35} subgroups; see Fig. 2b and Table S5. Subsequently, we calculated the relative change in disparity if these subgroups were to be omitted. We observe the greatest decrease in GDM disparity when the {Age ≥ 40} subgroup is omitted (Fig. 2c), which is reflected in the large decrease in GDM rate in Asian participants (Table S6).

3.2.2. Disparities in high-risk NewHTN subgroups—In NewHTN, the top three single-attribute subgroups with the largest LR^+ are {BMI ≥ 35}, {BMI = 30–35} and {Age = 35–39} (Fig. 3a), where the disparity measure is the highest in the {BMI = 25–30}, {BMI ≥ 35} and {Family DM = 1} subgroups (Table S3). The relative prevalence of NewHTN by race/ethnicity in each high-risk subgroup is highly heterogeneous: in high BMI groups such as {BMI = 25–30} and {BMI ≥ 35}, Asian participants have the highest rate of NewHTN, whereas White participants have the highest NewHTN rate in the {Age = 35–39} groups and Black participants have the highest NewHTN rate in the {Family DM = 1} group, as shown in Table S4.

When omitted from cohort, the top three single-attribute subgroups that result in the largest reduction in NewHTN rate were all BMI-related ({BMI ≥ 35}, {BMI = 30–35}, {BMI = 25–30}); see Fig. 3b. However, only the {BMI ≥ 35} subgroup led to a decrease in disparity measure when omitted (Fig. 3c). The racial/ethnic group for which the reduction in NewHTN risk was the highest was also different for each BMI subgroup, where omitting the {BMI ≥ 35} subgroup leads to the greatest reduction in NewHTN risk in Black participants, omitting the {BMI = 30–35} subgroup leads to the greatest reduction in NewHTN risk in Hispanic participants, and omitting the {BMI = 25–30} subgroup leads to the greatest reduction in NewHTN risk in Asian participants (Table S6).

3.2.3. Disparities in high-risk PReEc subgroups—The subgroup with the highest LR^+ for PReEc is the {High BP = 1} subgroup, followed by the {BMI ≥ 35} and {BMI = 30–35} subgroups (Fig. 4a, Table S3). The disparity measures for each of these subgroups are also similar, with {High BP = 1}, {PCOS = 1} and {Age < 18} being the three subgroups with the highest disparity (Fig. 4b), two of which are also in the highest disparity subgroups for PTB. The rates of PReEc by race/ethnicity are comparable as well, with Black participants having higher rates of PReEc across similar risk factors (Table S4).

The top three best PReEc risk-reducing when omitted single-attribute subgroups are {BMI ≥ 35}, {Diet = poor}, and {BMI = 30–35}. Among these high-risk subgroups, the {Diet = poor} subgroup is unique to PReEc and is not a high-risk subgroup found in other APOs in isolation (Table S5). The best disparity-reducing single-attribute subgroup when omitted is {High BP = 1}, followed by {BMI ≥ 35} and {Diet = poor}; see Table S5. The effect of omitting these subgroups on the overall rate of PReEc varied, where omitting the {High BP = 1} subgroup leads to the highest reduction in PReEc rate in Black participants, omitting {BMI ≥ 35} leads to significant reduction in both White and Black participants, and omitting the {BMI = 30–35} or {Family DM = 1} lead to the highest reduction in PReEc rate in Asian participants, although not statistically significant (Table S6).

3.2.4. Disparities in high-risk PTB subgroups—The landscape of disparity in PTB was vastly different from that in GDM. In PTB, the subgroup with the highest LR^+ is {High BP = 1}, followed by {Age < 18} and {PCOS = 1}; Fig. 5a. For these high-risk subgroups, the disparity measure is the highest in {Age < 18} followed by {Age = 35–39} and {High BP = 1}; see Table S3. The prevalence of PTB by racial/ethnic group also differed from that of GDM, with Black participants being the group with the highest PTB rate across high-risk

subgroups except those in the {Age < 18} subgroup, where the proportion of PTB patients are the highest among White participants (Table S4).

When omitting high-risk subgroups, we observe the greatest reduction in PTB rate is achieved when the {BMI ≥ 35} subgroup is omitted, followed by {Age < 18} and {High BP = 1} (Fig. 5b). Omitting the subgroup {BMI ≥ 35} led to highest reduction in disparity, followed by {Age < 18} and {High BP = 1}; see Table S5. Upon investigating the effect of omitting subgroup on PTB rate by race/ethnicity, we found all three high-risk subgroups where reduction in PTB prevalence is the most significant ({BMI ≥ 35}, {Age < 18}, {High BP = 1}) are also the groups that when omitted lead to the highest rate reduction in Black participants (Table S6).

3.3. Major APO risk-factors are associated with population structure

Given the frequent occurrence of Age and BMI as attributes in high-risk groups and the high variance in APO prevalence by race in these subgroups, we hypothesize that one of the components for disparities in APO could be partially attributable to the differences in age and BMI distributions between races/ethnicities in our cohort. We then employed the Kruskal-Wallis H-test on the age and BMI distributions marginalized by race and found the difference in distributions to be highly significant (Age: $p = 8.7 \times 10^{-280}$, BMI: $p = 7.0 \times 10^{-268}$); see Fig. 6.

4. Discussion

Adverse pregnancy outcomes can affect a family long after the delivery, and the ability to identify sources of disparities is crucial for ensuring equitable access to resources needed to address these outcomes. In this study, we used association rule mining as a tool to detect subgroups that are at increased risk for experiencing APOs, and evaluated the racial/ethnic disparities within these subgroups. We discovered significant heterogeneity in APO prevalence across racial/ethnic groups, quantified the disparity in each high-risk subgroup, and evaluated each subgroup's contribution to the total risk and disparity through observing the relative rate change when the subgroup was omitted from the cohort. In addition, we identified significant differences in age and BMI distributions across racial/ethnic groups, which appear to play an important role in shaping the APO risk landscape. The simplicity and interpretable nature of association rules also enable the findings to be accessible to wide audiences including clinicians and policy makers. While the study does not model clinical intervention, our findings can be used to inform planning of policy interventions, such as influencing resource allocation in communities where disparities and health outcomes need to be addressed. For example, the high prevalence of GDM among Asian participants above the age of 40 could serve as evidence for prioritizing education on the potential impact of maternal age on the risk of gestational diabetes in Asian communities, while the high prevalence of PReEc among Black participants with high blood pressure could serve as evidence for prioritizing education on blood pressure management in Black communities.

As with any clinical data, some variables used in our study may be underreported or incorrectly recorded. Additionally, the modest sample size resulted in relatively large confidence intervals in some high-risk subgroups. The change in APO proportion if a

subgroup is omitted also represents an idealized form of intervention with two strong assumptions; i.e., we assume that if an intervention on a risk factor is given, then (1) this risk factor is reduced to 0% in the population and (2) individuals who originally harbored these risk factors will proportionally distribute to other subgroups. These should not be taken as a realistic estimate of how much APO prevalence might decrease if an intervention is placed on a specific risk factor but rather an estimate of the contribution of the risk factors to the overall prevalence of APOs. It is also worth mentioning that when a high-risk subgroup is omitted but the disparity measure increases, it does not necessarily mean that addressing such a subgroup should not be performed; instead, it shows that some groups may not receive equal benefits from addressing these risk factors.

This study can be extended to include higher-resolution groupings of risk factors as well as the possibilities that other factors (e.g., social, economic, cultural) could have larger impact on disparities than the features investigated herein. Of note, however, this work does not provide evidence for biological differences between races and ethnicities that may predispose one over another towards certain APOs. Overall, this study calls for the investigation of disparities beyond the population level, and brings to attention the importance of considering subgroup-level disparities, which may manifest differently from their population form.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors acknowledge the support by the NIH grant R01HD101246 to DMH, SN, and PR.

References

1. Aungst R, Perspect Audiol 7, 29 (2011).
2. Lasser et al., Am J Public Health 96, 1300 (2006). [PubMed: 16735628]
3. Orsi et al., Am J Public Health 100, 349 (2010). [PubMed: 20019299]
4. Lavizzo-Mourney et al., N Engl J Med 384, 1681 (2021). [PubMed: 33951376]
5. Benjamins et al., JAMA Netw Open 4, e2032086 (2021). [PubMed: 33471116]
6. Laveist T and Gaskin D, Int J Health Serv 41, 231 (2011). [PubMed: 21563622]
7. Engelgau M et al., Ethn Dis 29, 103 (2019). [PubMed: 30906157]
8. Petersen E et al., Morb Mortal Wkly Rep 68, 762 (2019).
9. Grobman W et al., Obstet Gynecol 131, 328 (2018). [PubMed: 29324613]
10. Williams D and Rucker T, Health Care Financ Rev 21, 75 (2000). [PubMed: 11481746]
11. Haas D et al., Am J Obstet Gynecol 212, 539.e1 (2015).
12. Guerrero et al. , Genetic polymorphisms associated with adverse pregnancy outcomes in nulliparas, medRxiv 2022.02.28.22271641 (2022).
13. Artzi N et al., Nat Med 26, 71 (2020). [PubMed: 31932807]
14. Bao W et al., JAMA Intern Med 174, 1047 (2014). [PubMed: 24841449]
15. Pagel KA et al., JAMA Netw Open 5, e2229158 (2022). [PubMed: 36040739]
16. Agrawal R, Imieliński T and Swami A, ACM SIGMOD Rec 22, 207 (1993).
17. Tan PN, Steinbach M and Kumar V, Introduction to data mining (Pearson, 2006).

18. Glas AS et al., J Clin Epidemiol 56, 1129 (2003). [PubMed: 14615004]
19. Efron B and Tibshirani R, Stat Sci 1, 54 (1986).
20. Keppel K et al., Vital Health Stat 2, 1 (2005).
21. De Maio FG, J Epidemiol Community Health 61, 849 (2007). [PubMed: 17873219]
22. Benjamini Y and Hochberg Y, J R Stat Soc Series B 57, 289 (1995).
23. Lao T et al., Diabetes Care 29, 948 (2006). [PubMed: 16567851]
24. Dietl A and Farthmann J, Lancet 386, 1627 (2015).
25. Ferré C et al., Morb Mortal Wkly Rep 65, 1181 (2016).
26. Sheen J et al., Am J Obstet Gynecol 220, S222 (2019).
27. Makarem N et al., Circulation 145, A073 (2022).

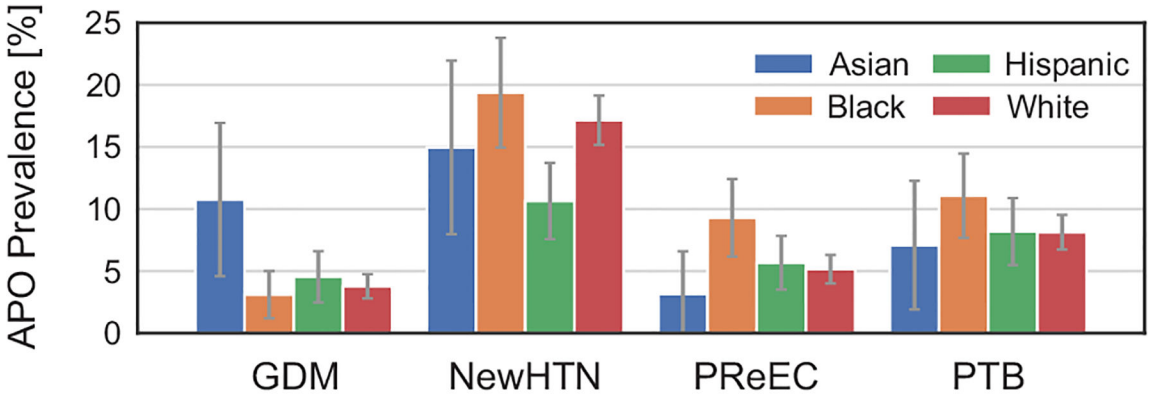

Fig. 1.

The prevalence of each adverse pregnancy outcome (APO) with respect to self-reported race/ethnicity. GDM: gestational diabetes mellitus; NewHTN: new hypertension; PReEc: preeclampsia; PTB: preterm birth. A pairwise comparison of APO rates by race/ethnicity is available in Table S2.

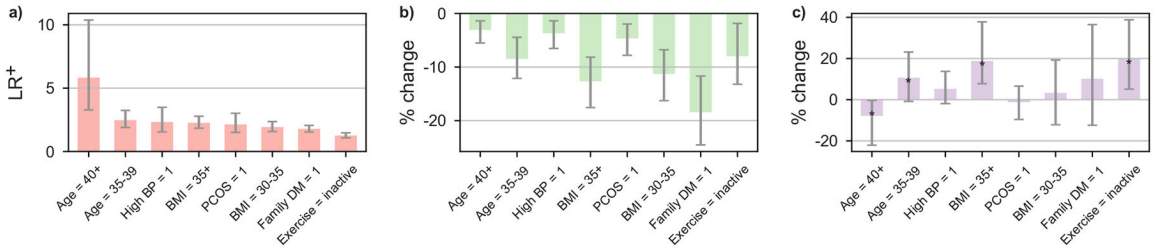

Fig. 2.

The prevalence and disparities of GDM and high-risk subgroup relative contribution to the disparity. (a) The LR^+ associated with high-risk GDM subgroups. (b) The relative change in GDM prevalence if a subgroup is omitted from the cohort. (c) The relative change in Gini coefficient if a subgroup is omitted from cohort, with markings for statistically significant values. Exact values and prevalence by each racial/ethnic group are available in Supplementary Tables S3–S6.

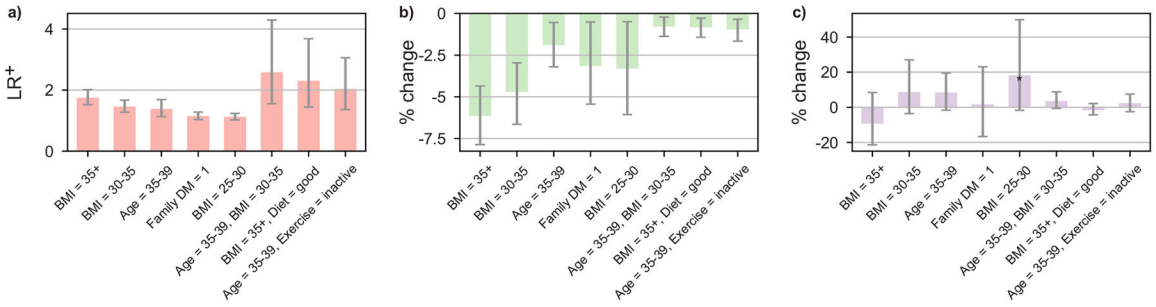

Fig. 3.

The prevalence and disparities of NewHTN and high-risk subgroup relative contribution to the disparity. (a) The LR^+ associated with high-risk NewHTN subgroups. (b) The relative change in NewHTN prevalence if a subgroup is omitted from the cohort. (c) The relative change in Gini coefficient if a subgroup is omitted from cohort, with markings for statistically significant values. Exact values and prevalence by each racial/ethnic group are available in Supplementary Tables S3–S6.

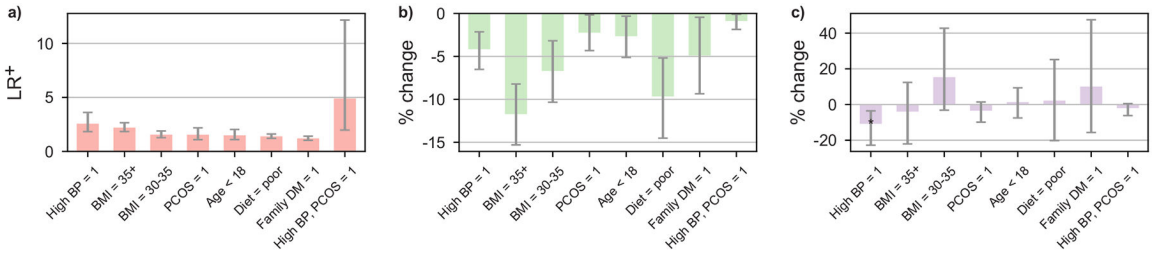

Fig. 4.

The prevalence and disparities of PReEc and high-risk subgroup relative contribution to the disparity. (a) The LR^+ associated with high-risk PReEc subgroups. (b) The relative change in PReEc prevalence if a subgroup is omitted from the cohort. (c) The relative change in Gini coefficient if a subgroup is omitted from cohort, with markings for statistically significant values. Exact values and prevalence by each racial/ethnic group are available in Supplementary Tables S3–S6.

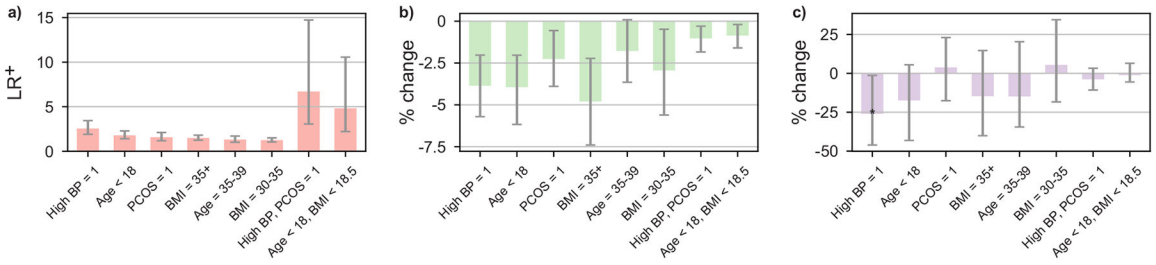

Fig. 5.

The prevalence and disparities of PTB and high-risk subgroup relative contribution to the disparity. (a) The LR^+ associated with high-risk PTB subgroups. (b) The relative change in PTB prevalence if a subgroup is omitted from the cohort. (c) The relative change in Gini coefficient if a subgroup is omitted from cohort, with markings for statistically significant values. Exact values and prevalence by each racial/ethnic group are available in Supplementary Tables S3–S6.

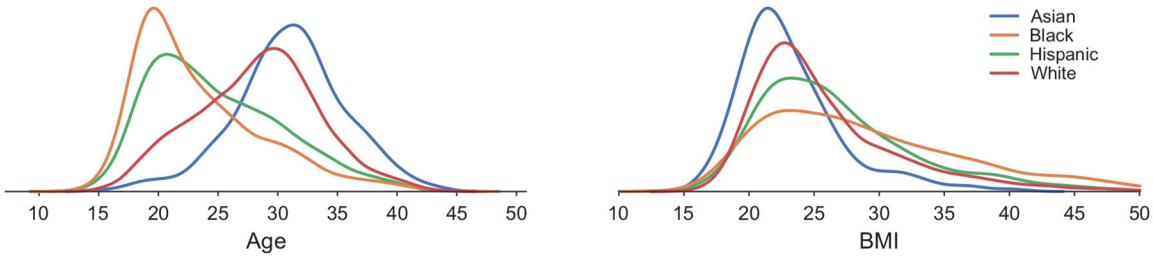

Fig. 6.

Age and BMI distributions for each racial/ethnic group in the cohort visualized using the kdeplot function from the Python library Seaborn.

Table 1.

Examples of statistically significant association rules for the nuMoM2b cohort.

Antecedent	Consequent	Confidence	LR^+ [95% CI]	Adjusted *p*-value
{Age = 35–39}	{GDM = 1}	9.6% (51/531)	2.5 [1.9, 3.2]	4.7×10^{-7}
{Age = 35–39}	{NewHTN = 1}	21.1% (112/531)	1.4 [1.1, 1.7]	1.1×10^{-2}
{Age < 18}	{PTB = 1}	14.3% (70/489)	1.8 [1.4, 2.3]	1.7×10^{-4}
{BMI ≥ 35}	{GDM = 1}	8.8% (78/882)	2.3 [1.8, 2.8]	5.8×10^{-9}
{BMI ≥ 35}	{PTB = 1}	11.9% (105/882)	2.2 [1.8, 2.7]	7.2×10^{-4}
{BMI ≥ 35}	{NewHTN = 1}	25.3% (223/882)	1.8 [1.5, 2.0]	1.1×10^{-10}
{Diet = poor}	{PReEc = 1}	7.9% (146/1853)	1.4 [1.2, 1.6]	3.0×10^{-4}
{Exercise = inactive, High BP = 1}	{PTB = 1}	21.4% (19/89)	2.9 [1.8, 4.8]	5.7×10^{-3}
{Diet = poor, High BP = 1}	{PReEc = 1}	22.2% (16/72)	4.7 [2.7, 8.1]	1.2×10^{-3}
{Age = 35–39, BMI = 30–35}	{NewHTN = 1}	33.3% (22/66)	2.6 [1.6, 4.3]	3.6×10^{-3}

Table 2.

Prevalence and count of APOs in each racial/ethnic group, their respective disparity measure and p-values from a chi-square (χ^2) test.

APO	Asian (n = 381)	Black (n = 1291)	Hispanic (n = 1587)	White (n = 5644)	Total	Gini	χ^2 p-value
GDM	41 (10.8%)	40 (3.1%)	72 (4.5%)	213 (3.8%)	366 (4.1%)	0.268	1.70×10^{-10}
NewHTN	57 (15.0%)	250 (19.4%)	169 (10.7%)	968 (17.2%)	1444 (16.2%)	0.114	9.32×10^{-11}
PReEc	12 (3.2%)	120 (9.3%)	90 (5.7%)	291 (5.2%)	513 (5.8%)	0.204	2.43×10^{-8}
PTB	27 (7.1%)	143 (11.1%)	130 (8.2%)	459 (8.1%)	759 (8.5%)	0.087	0.004