Differences in predicted rates of vaginal births after cesarean across racial groups in a ‘race-neutral’ model

Abstract

When physicians and pregnant patients make decisions about whether to pursue a vaginal birth or cesarean, there are many factors at play. While vaginal birth can have health benefits for both parent and child, there are significant safety risks. In order to minimize these risks, physicians use predictive models to determine how likely patients are to have successful vaginal births after cesareans (VBAC). For many years, these predictive models included race as a variable. This decision recently came under fire, and the Maternal Fetal Medicine Unit (MFMU) published a calculator that did not include race as a variable, but still predicted VBAC success with high accuracy. A large body of work in machine learning has highlighted that supposedly de-biased systems often re-code sensitive variables like race in terms of proxy variables. In order to determine if this was the case in this calculator, we replicated their formula, then found base-rate statistics of all the input variables for three different racial groups: Black, White, and Asian. We found that the distribution of VBAC probabilities for our simulated patients from these three groups were, indeed, significantly different from each other. Further, the predicted VBAC rates increased as a function of societal marginalization: Black patients were 47.6% likely to have a successful VBAC, Asian patients had a 48.6% probability, and White patients had a 49.4% probability. While these values are all within a few percentage points of each other, the differences in these simulated distributions shows how there may still be underlying disparities in the maternal healthcare system.

Introduction

The use of technology in the medical industry has progressively increased in recent years as a way to provide quicker, more efficient, and patient-specific care. Although very beneficial, introducing these technologies into an industry that has a history of discriminatory notions against America's marginalized groups could very easily impact the results in many of these advancements. The healthcare system has always been very apparent at how they are designed predominantly to benefit mostly white people, while hurting and disregarding people of color. This underrepresentation has resulted in both pre-existing racial disparities in medicine and those exacerbated by medical technology. In skin disease diagnostics, for example, the diagnostic criterias are based on symptoms prevalent in white patients, resulting in underdiagnosis for patients with darker skin (Ly et al., 2021). This is made worse through technology when algorithmic diagnostic tools are fed only canonical examples of skin disease.

In this paper, we investigate the Maternal-Fetal Medicine Units Network's (MFMU) vaginal birth after cesarean (VBAC) calculator, which predicts how likely a patient is to successfully give birth vaginally after a previous cesarean (Grobman et al., 2021). Initially, the MFMU calculator included a question that asked users for whether they were African American and/or Hispanic. However, after the network recognized there was potential for racial bias that they wanted to address, they removed race and ethnicity as variables, but still kept the accuracy rate of their predictions. So, we considered the fact that although efforts were made to eliminate any bias, the calculator could have still adopted proxy variables which were able to manipulate the results. This initiated our experiment and we questioned whether we could develop a model of the calculator without requiring the use of race and ethnicity as a variable and find possible explanations of where bias still arose. We started our experiment by replicating the VBAC calculator and then compiled a numerical dataset which included the averages of each variable: age, obstetric history, arrest of labor history, hypertension history, weight, and height for black, white, and asian women in America. We combined these two tools to create a patient simulator which filled in the variables for the calculator based on whichever race that was inputted. Then, we generated medical histories for 10,000 simulated patients and used the calculator to predict VBAC probability distributions and look for any differences in the probability likelihoods between each race.

Background

Initially, our project's scope was primarily on ethical computing and human rights; we then explored racism in the healthcare system before focusing on precision medicine and its flaws, which ultimately led to the investigation of the VBAC calculator. During our major focus on precision medicine, we examined three technical publications relating to the complexity of precision medicine in terms of how it handles race and ethnicity.

Long standing difficulties of prescribing medicine, due to its impreciseness have made physicians consider implementing variables to better personalize care given to patients. So, while race is not a biological component, it correlates with geographic heritage, a driver of genetic variety that might alter drug response, and thus became a common variable in precision medicine. However, this only oversimplified the complexity of utilizing race in this manner (Bonham et al., 2016). Prescribing medications like this tends to overlook the importance of ancestry, health disease, and drug response. It suggested that using precision medicine to replace race as a medical variable will only work if the data being used is representative of the community that should benefit from it. This furthered the conversation on how the healthcare system lacks a holistic representation in determining care for individuals. It also contributed to its claim by exhibiting different cases of precision medicine being faced with challenges pertaining to race and ethnicity to show us how a lot more needs to be done to perfect precision medicine before it can be used. For example, the first race-cased drug approved by the FDA, called BiDil, was used to treat heart failure in self-identified black patients. Some claimed that this was one step forward in revolutionizing personalized medicine, while others claimed that race as the only variable was not enough to be the determining factor of who gets to use the drug. This paper helped us in our research, as it gave insight into how precision medicine will never be a simple case and still needs a lot of work and significant change within the healthcare system to work.

In addition to understanding the ways that precision medicine exacerbates racial inequities, we wanted to learn more about how communities affected by racism perceive the use of predictive algorithms in medical care. A group of African American and Hispanic people were asked several questions pertaining to precision medicine to gather their perspectives on its benefits and disadvantages. The participants agreed that precision medicine is needed and could provide a lot of benefits but were also scared that the potential disadvantages racism could present might cause more problems (Yeh et al., 2020). The paper helped us in our research as it gave us a viewpoint of direct feedback and perspective on precision medicine from marginalized groups. But, because the participants came from different cities across the country, it's unclear if the differences are because of cultural beliefs, historical and social situations, or a combination of the two. This paper lets us understand how there will always be a certain amount of distrust from these groups with the medical industry due to the history they have with it.

From a broad perspective, it may appear that using factors such as race can perpetuate inequalities and reduce marginalized populations' trust and buy-in, however it can and does serve as a beneficial proxy when more granular data that would be useful is unavailable. Although it may seem unnecessary to use race/ethnicity as a variable in medicine, it could actually help capture trends for epidemiologic information. Race in the past has been used so much as a determiner for many factors today that there is no choice but to use it to look for these trends. In the past, racial injustice has compromised medicine by the exploitation of race and ethnicity as biological determinants, despite the reality that race and ethnicity have no biological basis. So, until better predictors are available, we should still use race/ethnicity to better address and eliminate health inequalities (Borrellet al., 2021). The paper contributed to our research as it also addressed the fact that the idea of race has become so complex and a major factor in our everyday lives, there is no point in being “color-blind”, but instead embracing it as a way to give us insight for other problems.

These papers were a crucial aspect to the start of our research as it was necessary that we kept all of this information in the back of our heads throughout the process. We felt that it was necessary to keep an open mind throughout our research project in order not to undermine the complexity of using such a powerful variable such as race in the healthcare system. Additionally, it motivated us to delve further into the potential sources of proxy variables and how they affect outcomes.

Dataset

Our numerical dataset for this project consisted of a compilation of datasets that were pulled from seven different sources. We had 6 variables we needed to get race-based distributions for:

Age Distributions:

Age	White	Black	Asian
Under 15 years	1018	571	18
15 years	2943	1311	46
16 years	7428	3030	123
17 years	14901	5679	200
18 years	29196	10375	418
19 years	52818	17772	809
20-24 years	475033	138212	13818
25-29 years	758281	172331	50562
30-34 years	802099	139808	93370
35-39 years	412488	75263	58594
40-44 years	85131	19035	12526
45-49 years	5534	1371	11120
50 years and over	560	221	185

We pulled age distributions for pregnant women from the US Census (2019 Population Estimates by Age, Sex, Race and Hispanic Origin).

Height and Weight:

Race	Height_Avg	Height_STD	Weight_Avg	Weight_STD
White	162.4	8.85742	92.2	26.53489401
Black	162.5	7.0852	90.8	27.64037084
Asian	156.3	10.00369	76.2	18.33670363

These two came from The National Bureau of Economic Research Vital Statistics Natality Birth Data. We were able to get both averages and standard deviations for each race.

Obstetric History:

Race	VBAC
White	0.3523015493
Black	0.2752468489
Asian	0.337235348

VBAC Rates: Racial and ethnic differences in the likelihood of vaginal birth after cesarean delivery

Race	1NUM	2aNUM	2bNUM	6NUM	1DEN	2aDEN	2bDEN	6DEN
White	53,457	58,344	72,310	33,408	460,331	237497	72310	34902
Black	11,709	12,174	16,558	3759	83803	40330	16,558	4196
Asian	7218	5223	10,939	3778	57324	19965	10939	3918

3	4a	4b
28,995/828874	29,142/391576	88,888/88888
9755/173022	8744/71433	23,179/23179
3048/79058	2228/22944	7907/7907

No vaginal history and No VBAC rate: [Veldes: Examining Cesarean Delivery Rates by Race: a Population-Based Analysis Using the Robson Ten-Group Classification System]
Rows 1, 2a, 2b, and 6 relate to No vaginal history
Rows 3, 4a, and 4b relate to No VBAC

Arrest of Labor History:

Race	No_Vag_History	No_VBAC
White	0.5666110065	0.08108744416
Black	0.4798186902	0.244934461
Asian	0.4710604511	0.191704201

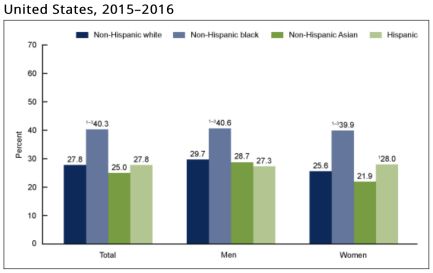
We got the 7.5% number from [Implementation Of New Definitions Of Labor Arrest Disorders And Failed Induction Can Decrease The Cesarean Rate], and the racial breakdown from [Racial and Ethnic Differences in Indication for Primary Cesarean Delivery at Term: Experience at One U.S. Institution] -- we then combined these two pieces of information to generate our race-based probabilities of arrest of labor, making sure they their population-normalized mean came out to 7.5% and that the differences in probability were correctly-proportioned

Hypertension:

Race	Hypertension_Likelihood
White	0.256000
Black	0.399000
Asian	0.219000

Pulled from the [National Center for Health Statistics Hypertension Prevalence and Control Among Adults: United States, 2015-2016 Data Brief]

Figure 2. Age-adjusted prevalence of hypertension among adults aged 18 and over, by sex and race and Hispanic origin: United States, 2015–2016



This graph from the paper is what we exclusively used for our data

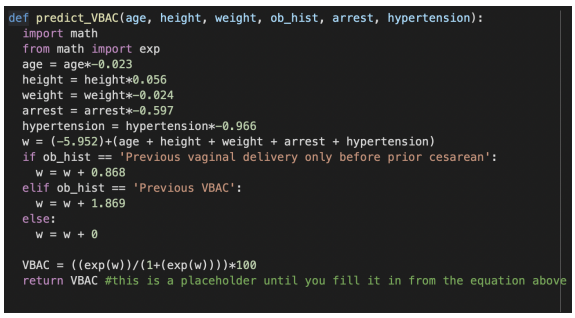
Methodology and Models

Our main goal was to determine whether there still is encoded racial bias in the MFMU's VBAC prediction calculator even though race and ethnicity are no longer included. In order to answer this question, we looked to examine if the anticipated VBAC probabilities for Black, White, and Asian patients would differ systematically. First, we had to replicate the MFMU VBAC calculator using the paper's linear model specified in their work (Grobman et al., 2021). Then, for each racial group, we identified the baseline distributions of the six input variables (age, height, weight, obstetric history, history of labor interruption, and hypertension). Using these distributions, we generated ten thousand simulated patients representing each of our three racial groups, forecast the likelihood that they would undergo a VBAC, and then compared the probability distributions to one another to check for any significant differences.

We were able to start the first step of our process of defining a replicate calculator in our code by using the weighted equation given by the MFMU network and named it "predict_vbac". Our calculator provides a probability score of the estimated likelihood of a vaginal birth following a cesarean depending on user inputs and has six parameters accounting for the six variables (questions) required by the calculator. Three of the variables require numbers as inputs within a defined range, these include: age (15-50 years old), height (119-191 cm), and pre-pregnancy weight (34-206 kg). Two of the variables are questions which require either a yes or no input, they are: arrest disorder indication for prior cesarean? and treated chronic hypertension? The last variable is for obstetrics history that requires one of three possible inputs: previous VBAC, no previous vaginal history, or previous vaginal delivery only before prior cesarean. After creating our calculator we tested its accuracy compared to the MFMU calculator in three different test cases by seeing the difference between both probability scores after entering the same random inputs for each variable. On average our calculator's scores were off by only 0.5% of the MFMU's calculator for all three test cases.

Using all individuals in the analytical cohort, the regression was as follows: predicted probability of VBAC $= (\exp(w)/[1+\exp(w)]) \times 100$, where $w = -5.952 - 0.023$ (age) $- 0.024$ (pre-pregnancy weight, kg) $+ 0.056$ (height, cm) $- 0.597$ (arrest indication) $+ 0.868$ (previous vaginal delivery only before prior cesarean) $+ 1.869$ (previous VBAC) $- 0.966$ (treated chronic hypertension) and with arrest indication, previous vaginal delivery only before previous cesarean delivery, previous VBAC, and treated chronic hypertension coded as 0 for "no" and 1 for "yes." This

The equation we used to replicate the MFMU calculator was given to us on page 5 of the network's paper (Grobman et al., 2021).

	The equation implemented in our code to define the predict_VBAC function.

The next step in our process was to create our simulated patient generator which we called, “get_simulated_patient”. We had to first compile a large dataset with data derived from multiple sources which contained: the average heights and its standard deviation, average weights and its standard deviation, arrest likelihoods, hypertension likelihoods, age distribution for pregnancies, and averages of the three inputs for ob history; the variables were each separated by race. We then imported these data back into our code. Using this we were able to create a randomized representative sample for each of the variables based on the three races. Height and weight were both safely assumed to have a normal distribution, as they were the only ones that contained averages along with standard deviations, so we used the “numpy.random.normal()” function to get our representative samples. Arrest likelihoods, hypertension likelihoods, and ob history were categorical variables, determined by probability of occurrence. For these variables we used the function, “numpy.random.random()” which works essentially like a weighted coin; the function generates a random number between 0 and 1 and if that number lands below the likelihood of that event happening, the function is true. Finally, for age distribution, the data we were using was already a representative sample, so we implemented the “numpy.cumsum()” function and used the “numpy.random.random()” function once again to generate a random age. The “get_simulated_patient” function returns a randomized representative sample for all six variables.

Adding our dataset into our code to create representative samples for the patient simulator:

```
import numpy as np
import re
import random
import statistics

def get_simulated_patient(Race):
  demo = demographics[demographics['Race'] == Race]

  #First get the two variables defined by normal distributions
  height = np.random.normal(demo['Height_Avg'], demo['Height_STD'])[0]
  weight = np.random.normal(demo['Weight_Avg'], demo['Weight_STD'])[0]

  #Then get the variables determined by likelihood
  if np.random.random() < demo['Hypertension_Likelihood'].astype('float').iloc[0]:
    hypertension = True
  else:
    hypertension = False

  ##OB_HIST
  ob_prob = np.random.random()
  if ob_prob < demo['No_Vag_History'].astype('float').iloc[0]:
    ob_hist = "No_Vag_History"
  elif ob_prob < (demo['VBAC'].astype('float').iloc[0]) + (demo['No_Vag_History'].astype('float').iloc[0]):
    ob_hist = "VBAC"
  else:
    ob_hist = "No_VBAC"
  ##OB_HIST

  if np.random.random() < demo['Arrest_Likelihood'].astype('float').iloc[0]:
    arrest = True
  else:
    arrest = False

  #Now let's get the variable where we have a defined distribution: age
  arr = np.array(ages[Race])
  int_arr = np.array(list(map(int, arr)))
  cs_arr = np.cumsum(int_arr/sum(int_arr))

  age_prob = np.random.random()
  for i in range(len(cs_arr)):
    if cs_arr[i] > age_prob:
      break
  age_range = (ages['Age'][i])
  if "-" not in age_range:
    new_result = re.findall('[0-9]+', age_range)
    age = (int(new_result[0]))
  else:
    new_result = re.findall('[0-9]+', age_range)
    age = (random.randint(int(new_result[0]), int(new_result[1])))

  return (age, height, weight, ob_hist, arrest, hypertension)
```

For our last step we used the “get_simulated_patient” along with the “predict_VBAC” function to create 10,000 randomized patient histories for each race and output their VBAC scores, the we used matplotlib’s plt.hist() function to plot a histogram for all the scores. This was done to check for any differences in the overall distribution of the graphs and their averages. Then we conducted a t-test for independence and checked if there were any significant differences among the mean probabilities.

Code for our last step:

```
#Asian Patient
VBAC_prob_Asian = []
for i in range(0, 10000):
  age, height, weight, ob_hist, arrest, hypertension = get_simulated_patient('Asian')
  VBAC_prob_Asian.append((predict_VBAC(age, height, weight, ob_hist, arrest, hypertension)))

x = VBAC_prob_Asian
plt.xlim((0,100))
plt.ylim((0,300))
plt.hist(x,bins = 100)
plt.xlabel('predicted probability of VBAC')
plt.ylabel('number of simulated patients')
print('mean:', sum(x)/10000)
print('median:', statistics.median(x))
print('standard deviation:', np.std(x))
```

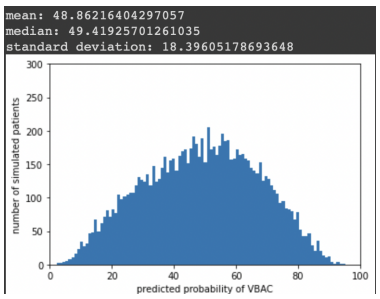
```
#White Patient
VBAC_prob_White = []
for i in range(0, 10000):
  age, height, weight, ob_hist, arrest, hypertension = get_simulated_patient('White')
  VBAC_prob_White.append((predict_VBAC(age, height, weight, ob_hist, arrest, hypertension)))

x = VBAC_prob_White
plt.xlim((0,100))
plt.ylim((0,300))
plt.hist(x,bins = 100)
plt.xlabel('predicted probability of VBAC')
plt.ylabel('number of simulated patients')
print('mean:', sum(x)/10000)
print('median:', statistics.median(x))
print('standard deviation:', np.std(x))
```

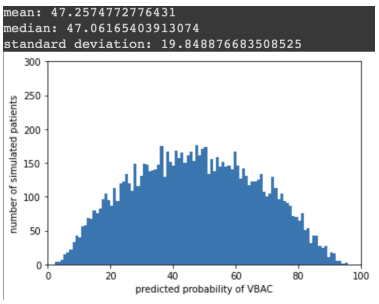
```
#Black Patient
VBAC_prob_Black = []
for i in range(0, 10000):
  age, height, weight, ob_hist, arrest, hypertension = get_simulated_patient('Black')
  VBAC_prob_Black.append((predict_VBAC(age, height, weight, ob_hist, arrest, hypertension)))

x = VBAC_prob_Black
plt.xlim((0,100))
plt.ylim((0,300))
plt.hist(x,bins = 100)
plt.xlabel('predicted probability of VBAC')
plt.ylabel('number of simulated patients')
print('mean:', sum(x)/10000)
print('median:', statistics.median(x))
print('standard deviation:', np.std(x))
```

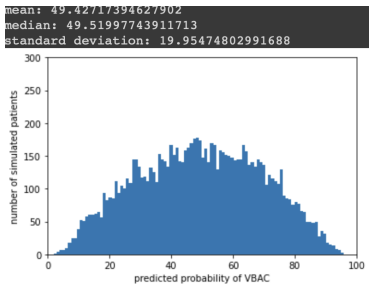
Results and Discussion

By eye, the distributions of VBAC probabilities for Asian, Black, and White patients look quite similar. However, we later conducted three pairwise t-tests comparing each patient group to each other. After Bonferonni correction for multiple hypothesis testing, the distributions of Black-White and Black-Asian VBAC scores were significantly different from each other (p-values of 4e-14 and 9e-09 respectively). The differences in the means between each of the populations were minimal, staying within the 1-2% range. The fact that the distributions of predicted VBAC scores were different for different racial groups suggests that the original calculator might have re-coded race in terms of other variables. Moving forward, it would likely be useful to build a linear model that predicts a patient's race from their biological variables for our simulated patients in order to determine which biological variables are most likely to be a proxy for race.

Mean, median, and standard deviation VBAC probability results for 10,000 simulates **Asian** Patients	Mean, median, and standard deviation VBAC probability results for 10,000 simulates **Black** Patients	Mean, median, and standard deviation VBAC probability results for 10,000 simulates **White** Patients







Code for the three pairwise t-tests:
1. **Black-Asian** VBAC scores
2. **Asian-White** VBAC scores
3. **White-Black** VBAC scores

```
import scipy.stats
print(scipy.stats.ttest_ind(VBAC_prob_Asian, VBAC_prob_Black))
print(scipy.stats.ttest_ind(VBAC_prob_Asian, VBAC_prob_White))
print(scipy.stats.ttest_ind(VBAC_prob_White, VBAC_prob_Black))
```

```
Ttest_indResult(statistic=5.929206460996766, pvalue=3.0940467475280947e-09)
Ttest_indResult(statistic=-2.0816921989969908, pvalue=0.03738328001843476)
Ttest_indResult(statistic=7.708469263810287, pvalue=1.3326666011685186e-14)
```

Conclusion

The main motivation of our research was to understand the impact of proxy variables hidden within the VBAC calculator. The results from our t-test tells us that the difference was not due to chance or a sampling error, but that there still remain underlying factors that are the cause of these 1-2% differences in the averages. These factors can range from under representative samples to the overall healthcare given to black people in America. The next steps that should be taken with this research is to find datasets that are more representative of all three races including ethnicity and to determine which biological variables proxy race. In consideration to the overall usage of similar algorithms in the healthcare system, it would be beneficial if we began to consider ethical computing as a prerequisite to the design processes of these advancements and be able to identify how and why bias can infiltrate in to the makeup of these algorithms. We can also attempt to dive further into a more non-medical and public health setting to understand how the social determinants of health (SDOH) have been able to impact such variables. Implementing such research before creating medical technology should be done as a way to educate ourselves about our community and understand how marginalized groups are disproportionately disadvantageous in almost every aspect of the SDOH.

References

1. Bonham, Vence L et al. "Will Precision Medicine Move Us beyond Race?." *The New England journal of medicine* vol. 374,21 (2016): 2003-5. doi:10.1056/NEJMp1511294

2. Borrell, Luisa N et al. "Race and Genetic Ancestry in Medicine - A Time for Reckoning with Racism." *The New England journal of medicine* vol. 384,5 (2021): 474-480. doi:10.1056/NEJMms2029562

3. Grobman, William A et al. "Prediction of vaginal birth after cesarean delivery in term gestations: a calculator without race and ethnicity." *American journal of obstetrics and gynecology* vol. 225,6 (2021): 664.e1-664.e7. doi:10.1016/j.ajog.2021.05.021

4. Ly, Dan P et al. "Black-White Differences in the Clinical Manifestations and Timing of Initial Lyme Disease Diagnoses - Journal of General Internal Medicine." SpringerLink, Springer International Publishing, 30 Sept. 2021, doi.org/10.1007/s11606-021-07129-1

5. Yeh, Vivian M et al. "Can Precision Medicine Actually Help People Like Me? African American and Hispanic Perspectives on the Benefits and Barriers of Precision Medicine." Ethnicity & disease vol. 30,Suppl 1 149-158. 2 Apr. 2020, doi:10.18865/ed.30.S1.149