Analysis of Disparity Between Scleroderma Clinical Trial Participant Demographics and Scleroderma Patient Demographics in the United States

Natalie Goulett

2024-09-12

# 1. Summary

To investigate whether the demographics of scleroderma clinical trial participants are reflective of the scleroderma patient population, we conduct a Pearson’s chi-square test to compare the proportion of Black patients enrolled in scleroderma clinical trials to the estimated proportion of black patients with scleroderma in the United States. Furthermore, we will quantify this disparity by estimating the percent of all Black scleroderma patients that are unrepresented in clinical trials in the United States. We will focus specifically on the southeastern U.S., which is home to about half of the United States’ total black population.

## 1.1 Data Sources

We suspect the demographics of scleroderma patients in the United States are similar to those of the patient population described in a [census by Mayes, et al., 2003](https://pubmed.ncbi.nlm.nih.gov/12905479/). This census measured the prevalence of systemic scleroderma in a patient population in Detroit, MI in 2003. We will use prevalence estimates provided by this study to estimate the actual prevalence of scleroderma among the black population in the United States.

Our U.S. clinical trial demographic data were gathered from all publicly available NCBI clinical trials on scleroderma conducted in the United States that were available on [ClinicalTrials.gov](https://clinicaltrials.gov/). Population estimates and demographics for the U.S. and each U.S. state as of July 30th, 2024 were provided by the [U.S. Census Bureau](https://www.census.gov/quickfacts/fact/table/US/PST045222).

## 1.2 Methods

To conduct the Pearson’s chi-square test on the proportion of Black patients enrolled in scleroderma clinical trials to the estimated proportion of black patients with scleroderma in the United States, we will compare the proportion of Black participants from the population studied in Mayes, et al. to the proportion of Black patients that have participated in publicly available U.S. clinical trials on scleroderma as of July 30th, 2024.

# 2. U.S. Clinical Trial vs. Scleroderma Patient Demographics Chi-Square Test

## 2.1 Packages Used

# Install and load required packages:  
# install.packages("tidyverse")  
# install.packages("gt")  
  
library(tidyverse) # for data wrangling and visualization  
library(gt) # for creating table

## 2.2 Data Transformation

ct\_demo\_df <- read\_csv(  
 "./Data/20240717\_usSScClinTrialDemo.csv",  
 col\_names = TRUE,  
 na = c("#N/A", "#NA", "Not Reported", "Not reported")) %>%   
 rename(  
 "total\_participants" = "Number of participants analyzed - use this number for total participants",  
 "female" = "Female",  
 "male" = "Male",  
 "other\_gender" = "Other",  
 "study\_ID" = "ClinicalTrials.gov ID",  
 "hispanic" = "Hispanic or Latino",  
 "non\_hispanic" = "Not Hispanic or Latino",  
 "unknown\_ethnicity" = "Unknown or not reported ethnicity",  
 "native\_american\_an" = "American Indian or Alaska Native",  
 "asian" = "Asian",  
 "native\_hawaiian\_pi" = "Native Hawaiian or Other Pacific Islander",  
 "black\_participants" = "Black or African American",  
 "white\_participants" = "White",  
 "multi\_race" = "More than one race",  
 "unknown\_race" = "Unknown or not reported"  
 ) %>%   
 drop\_na(total\_participants, black\_participants)

## 2.3 Pariticpant Demographics Calculation

library(dplyr)  
  
# total U.S. clinical trial participants  
ct\_n <- sum(ct\_demo\_df$total\_participants, na.rm = TRUE)  
  
# Summarize participants by race, gender, and ethnicity  
race\_summary <- ct\_demo\_df %>%  
 summarise(  
 total\_black = sum(black\_participants, na.rm = TRUE),  
 total\_white = sum(white\_participants, na.rm = TRUE),  
 total\_native\_american\_an = sum(native\_american\_an, na.rm = TRUE),  
 total\_native\_hawaiian\_pi = sum(native\_hawaiian\_pi, na.rm = TRUE),  
 total\_asian = sum(asian, na.rm = TRUE),  
 total\_multi\_race = sum(multi\_race, na.rm = TRUE),  
 total\_unknown\_race = sum(unknown\_race, na.rm = TRUE),  
 percent\_black = (total\_black / ct\_n) \* 100,  
 percent\_white = (total\_white / ct\_n) \* 100,  
 percent\_native\_american\_an = (total\_native\_american\_an / ct\_n) \* 100,  
 percent\_native\_hawaiian\_pi = (total\_native\_hawaiian\_pi / ct\_n) \* 100,  
 percent\_asian = (total\_asian / ct\_n) \* 100,  
 percent\_multi\_race = (total\_multi\_race / ct\_n) \* 100,  
 percent\_unknown\_race = (total\_unknown\_race / ct\_n) \* 100,  
 total\_reporting\_race = sum(  
 total\_black,  
 total\_white,  
 total\_native\_american\_an,  
 total\_native\_hawaiian\_pi,  
 total\_asian,  
 total\_multi\_race  
 )  
 )  
  
gender\_summary <- ct\_demo\_df %>%   
 summarise(  
 total\_male = sum(male, na.rm = TRUE),  
 total\_female = sum(female, na.rm = TRUE),  
 percent\_male = (total\_male / ct\_n) \* 100,  
 percent\_female = (total\_female / ct\_n) \* 100  
 )  
  
ethnicity\_summary <- ct\_demo\_df %>%  
 summarise(  
 total\_hispanic = sum(hispanic, na.rm = TRUE),  
 total\_non\_hispanic = sum(non\_hispanic, na.rm = TRUE),  
 total\_unknown\_ethnicity = sum(unknown\_ethnicity, na.rm = TRUE),  
 total\_reporting\_ethnicity = sum(total\_hispanic, total\_non\_hispanic),  
 percent\_hispanic = (total\_hispanic / total\_reporting\_ethnicity) \* 100,  
 percent\_non\_hispanic = (total\_non\_hispanic / total\_reporting\_ethnicity) \* 100  
 )  
  
# Combine summaries into one data frame  
demographics\_df <- bind\_cols(  
 race\_summary,  
 gender\_summary,  
 ethnicity\_summary  
) %>%   
 pivot\_longer(cols = everything())  
  
print(demographics\_df, n = 25)

# A tibble: 25 × 2  
 name value  
 <chr> <dbl>  
 1 total\_black 60   
 2 total\_white 570   
 3 total\_native\_american\_an 1   
 4 total\_native\_hawaiian\_pi 2   
 5 total\_asian 15   
 6 total\_multi\_race 25   
 7 total\_unknown\_race 18   
 8 percent\_black 8.68   
 9 percent\_white 82.5   
10 percent\_native\_american\_an 0.145  
11 percent\_native\_hawaiian\_pi 0.289  
12 percent\_asian 2.17   
13 percent\_multi\_race 3.62   
14 percent\_unknown\_race 2.60   
15 total\_reporting\_race 673   
16 total\_male 111   
17 total\_female 580   
18 percent\_male 16.1   
19 percent\_female 83.9   
20 total\_hispanic 31   
21 total\_non\_hispanic 405   
22 total\_unknown\_ethnicity 63   
23 total\_reporting\_ethnicity 436   
24 percent\_hispanic 7.11   
25 percent\_non\_hispanic 92.9

## 2.4 Create Contingency Table

# calculate number of black U.S. clinical trial participants  
ct\_n\_black <- sum(ct\_demo\_df$black\_participants, na.rm = TRUE)  
# calculate number of non-black U.S. clinical trial participants  
ct\_n\_nonblack <- ct\_n - ct\_n\_black  
  
# Input demographics of patient population characterized in Mayes et al., 2003  
mayes\_n <- 706  
mayes\_n\_black <- 186  
mayes\_n\_nonblack <- mayes\_n - mayes\_n\_black  
  
# Create contingency table of expected and observed demographics  
demo\_contingency\_table <- matrix(  
 c(  
 ct\_n\_black,  
 mayes\_n\_black,  
 ct\_n\_nonblack,  
 mayes\_n\_nonblack  
 ),  
 nrow = 2  
)  
  
colnames(demo\_contingency\_table) <- c("Black", "Non-Black")  
rownames(demo\_contingency\_table) <- c("Clinical Trial Participants", "Scleroderma Patients")

## 2.5 Pearson’s Chi-Square Test on Demographics

demo\_chi\_sq <- chisq.test(  
 demo\_contingency\_table,  
 correct = TRUE  
)  
  
print(demo\_chi\_sq)

Pearson's Chi-squared test with Yates' continuity correction  
  
data: demo\_contingency\_table  
X-squared = 73.876, df = 1, p-value < 2.2e-16

### 2.5.1 Interpretation

The proportion of black scleroderma patients studied in Mayes et al. is significantly larger than the proportion of black participants enrolled in scleroderma clinical trials available on [ClinicalTrials.gov](https://clinicaltrials.gov/), (1, N = 1397) = 73.9, *p* < .001. Thus, it is suspected that the true proportion of black scleroderma patients in the U.S. is greater than that in U.S. clinical trials.

# 3. Black Representation in U.S. Scleroderma Clinical Trials

The number of black scleroderma patients in the U.S. and southeastern U.S. states that are unrepresented in U.S. clinical trials was estimated using prevalence data from [Mayes, et al., 2003](https://pubmed.ncbi.nlm.nih.gov/12905479/) and demographic data and population size estimates from the [U.S. Census Bureau](https://www.census.gov/quickfacts/fact/table/US/PST045222).

## 3.1 Representation Estimates Calculation

# Create data frame of state demographics from US Census Bureau estimates  
  
# Estimate scleroderma prevalence in U.S. black population from Mayes et al.  
scPrevalence\_Black <- 315.1/1000000  
# Estimate scleroderma prevalence in total U.S. population from Mayes et al.  
scPrevalence\_Total <- 242/1000000  
# Percent of black patients in U.S. scleroderma clinical trials  
us\_ct\_PercentBlack <- ct\_n\_black/ct\_n  
  
# Read in state population and percent black data from U.S. Census Bureau  
statePop\_df <- read\_csv("Data/20240812\_SeStatePopulations.csv")  
  
# Create data frame of Southeast states  
# Exclude the row representing the entire U.S.  
southeast\_df <- statePop\_df[statePop\_df$state != "United\_States", ]  
  
# Calculate the total population for the southeastern states  
se\_population <- sum(southeast\_df$population)  
  
# Calculate the weighted average of the percent black population  
weighted\_percent\_black <- sum(  
 southeast\_df$population \* southeast\_df$percentBlack  
 ) / se\_population  
  
# Create a new row for the southeastern U.S.  
southeast\_row <- data.frame(  
 state = "Southeastern US",  
 population = se\_population,  
 percentBlack = weighted\_percent\_black  
)  
  
# Append the new row to the original data.frame  
statePop\_df <- rbind(statePop\_df, southeast\_row)  
  
# Calculate percent of unrepresented black patients for each region and state  
stateScRep <- statePop\_df %>%   
 mutate(  
 blackPop = population \* percentBlack,  
 blackPatients = blackPop \* scPrevalence\_Black,  
 totalPatients = population \* scPrevalence\_Total,  
# Calculate represented black patients; 8.9% of the total patient population  
 repBlackPatients = totalPatients \* us\_ct\_PercentBlack,  
# Calculate unrepresented black patients by subtracting represented black  
# patients from total black patients  
 unrepBlackPatients = blackPatients - repBlackPatients,  
 percentUnrep = unrepBlackPatients / blackPatients,  
# Calculate the proportion of scleroderma patients that identify as black  
 proportionBlack = blackPatients / totalPatients  
 ) %>%   
 arrange(desc(population))

# Create representation by state table  
stateScRep %>%   
 select(  
 state,  
 percentBlack,  
 blackPatients,  
 totalPatients,  
 proportionBlack,  
 unrepBlackPatients,  
 percentUnrep  
 ) %>%   
 mutate(state = gsub("\_", " ", state)) %>%  
 gt() %>%   
 cols\_label(  
 state = "",  
 percentBlack = "Percent of Population Identifying as Black",  
 blackPatients = "Estimated Black Scleroderma Patients",  
 totalPatients = "Estimated Total Scleroderma Patients",  
 proportionBlack = "Percent of Scleroderma Patients who are Black",  
 unrepBlackPatients = "Estimated Unrepresented Black Patients",  
 percentUnrep = "Percent of Patients Unrepresented"  
 ) %>%   
 fmt\_number(  
 columns = c(blackPatients, totalPatients, unrepBlackPatients),  
 decimals = 0,  
 drop\_trailing\_dec\_mark = TRUE,  
 use\_seps = TRUE,  
 sep\_mark = ","  
 ) %>%   
 fmt\_percent(  
 columns = c(percentBlack, proportionBlack, percentUnrep),  
 decimals = 1  
 ) %>%   
 tab\_header(  
 title = "Black Representation in U.S. Scleroderma Clinical Trials by State"  
 ) %>%   
 tab\_source\_note(  
 source\_note = md(  
 "\*Note.\* National and state population data were provided by the U.S. Census Bureau. U.S. clinical trial demographics were sourced from publicly available NCBI clinical trial data available on ClinicalTrials.gov."  
 )  
 ) %>%   
 opt\_table\_font(  
 font = "Times New Roman"  
 )

Table 1: Black Representation in U.S. Scleroderma Clinical Trials by State

|  | Percent of Population Identifying as Black | Estimated Black Scleroderma Patients | Estimated Total Scleroderma Patients | Percent of Scleroderma Patients who are Black | Estimated Unrepresented Black Patients | Percent of Patients Unrepresented |
| --- | --- | --- | --- | --- | --- | --- |
| United States | 13.7% | 14,458 | 81,049 | 17.8% | 7,420 | 51.3% |
| Southeastern US | 22.4% | 5,448 | 18,680 | 29.2% | 3,826 | 70.2% |
| Florida | 16.9% | 1,204 | 5,472 | 22.0% | 729 | 60.5% |
| Georgia | 33.2% | 1,154 | 2,669 | 43.2% | 922 | 79.9% |
| North Carolina | 22.1% | 755 | 2,622 | 28.8% | 527 | 69.8% |
| Tennessee | 16.5% | 371 | 1,725 | 21.5% | 221 | 59.6% |
| South Carolina | 26.0% | 440 | 1,300 | 33.9% | 327 | 74.4% |
| Alabama | 26.6% | 428 | 1,236 | 34.6% | 321 | 74.9% |
| Louisiana | 32.6% | 470 | 1,107 | 42.4% | 374 | 79.5% |
| Kentucky | 8.8% | 126 | 1,095 | 11.5% | 30 | 24.2% |
| Arkansas | 15.6% | 151 | 742 | 20.3% | 86 | 57.3% |
| Mississippi | 37.8% | 350 | 711 | 49.2% | 288 | 82.4% |
| *Note.* National and state population data were provided by the U.S. Census Bureau. U.S. clinical trial demographics were sourced from publicly available NCBI clinical trial data available on ClinicalTrials.gov. | | | | | | |

## 3.2 Interpretation

Of the 14,458 estimated black scleroderma patients in the United States, about 7,420 (51.3%) are unrepresented in clinical trials. Approximately 3,826 of these unrepresented patients reside in the southeast region of the United States, where an estimated 70.4% of black scleroderma patients are unrepresented in clinical trials. Between the southeastern states, the estimated percent of unrepresented black scleroderma patients ranges from 24.2% in Kentucky to 82.4% in Mississippi.