Introduction

The data set selected was found on the Harvard Dataverse, in the Medicine, Health and Life Sciences section. The data set is called '**The impact of socioeconomic status on individual attitudes and experience with clinical trials: How socioeconomically disadvantaged individuals are being left behind'.** It was published October 10, 2023, by Jennifer Kim (Tufts University). It is a survey about people's attitudes toward clinical trials and experience being asked to participate in clinical research. There are 4,006 subjects that responded to this study.

The research question being asked is: Are socioeconomically disadvantaged individuals less likely to participate in clinical research? Socioeconomically disadvantaged people are defined as having lower levels of education and lower rates of income.

Methods

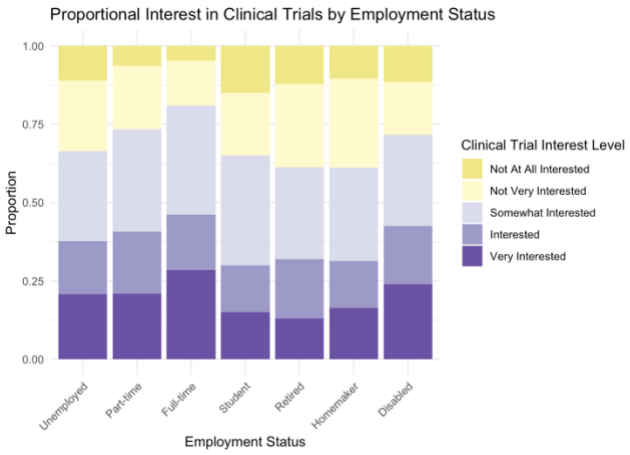
The data set was sourced from the Harvard Dataverse as a .csv file, which was subsequently converted into an Excel workbook for initial cleaning and preparation. In Excel, I simplified complex variable names, such as manually changing the 'Other - Write In (Required):Why aren't you interested in joining a clinical research study? [Select all that apply]' column to 'Q2_other_text.' Furthermore, all character responses of 'Yes' were uniformly coded as 1, while 'No' responses were coded as 0 to enhance data consistency. Columns were reorganized to optimize data relevance for this research project, moving less relevant variables to the end. The refined Excel file was then imported into the R programming environment using the 'read_excel' function for more thorough cleaning and data manipulation.

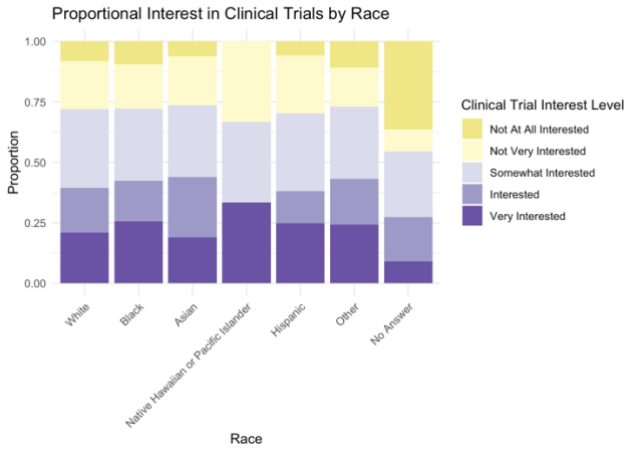
All relevant variables were converted from character strings to numeric categories. Variables that were previously distributed across multiple columns, like 'Race,' were consolidated into a single column, 'total_race,' containing distinct race categories. The data was re-labeled with more accurate and comprehensible descriptors for the purpose of visualization and analysis. A focused subset of the original data set was created, including key variables of interest: Gender, State, Age, Interest in Participating in a Clinical Trial (labeled as 'interest'), Years of Work Experience (WorkExp), Household Size (Household), Number of Dependents (Dependents), Employment Status (EmpStat), Race (total_race), Annual Income (AnnualIncome), Education Level (Ed), Living Situation (LivingSit), English Proficiency Level (EnglishProf), Help with Understanding Medical Literature (HelpMedLit), Location Type (Location: city, suburb, rural), Insurance Status (Insurance), Perceived Benefits of Participating in a Trial (Benefit), and Perceived Risks of Participation (Risk). There was no missing data, and the final analyzed data set included all 4,006 subjects from 51 US states.

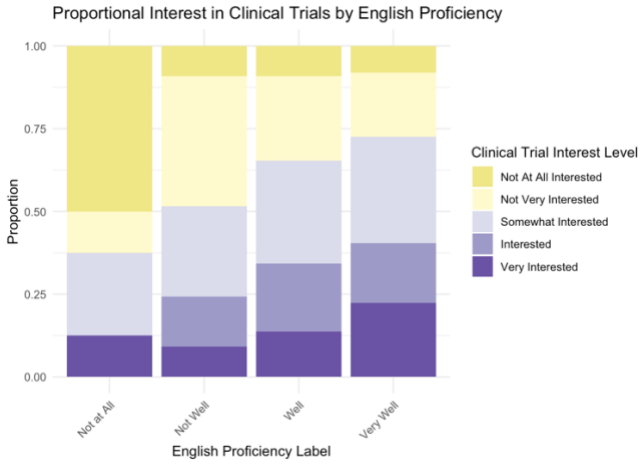
Bar plots were initially generated for each key variable of interest to visualize their distributions and identify any outliers, missing values, or meaningful patterns. Subsequently, stacked bar plots were constructed to illustrate the counts of groups within each key variable, color-coded by interest in clinical trials.

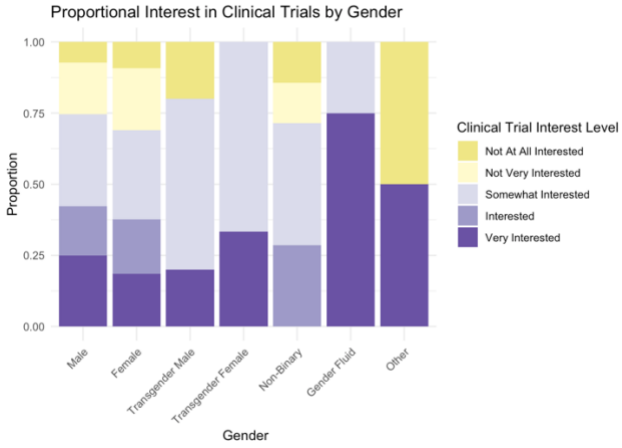
To facilitate a more intuitive understanding of the data, these same plots were recreated with counts displayed proportionally. Additionally, heatmaps were generated to examine the differences in perceived benefits and risks across key variables, based on patterns related to clinical trial interest. Stacked bar charts were created as supplementary visualizations for the respective heatmaps. Finally, maps were produced to explore any discernible patterns between the average interest level and average income by state, considering the primary variables of interest.

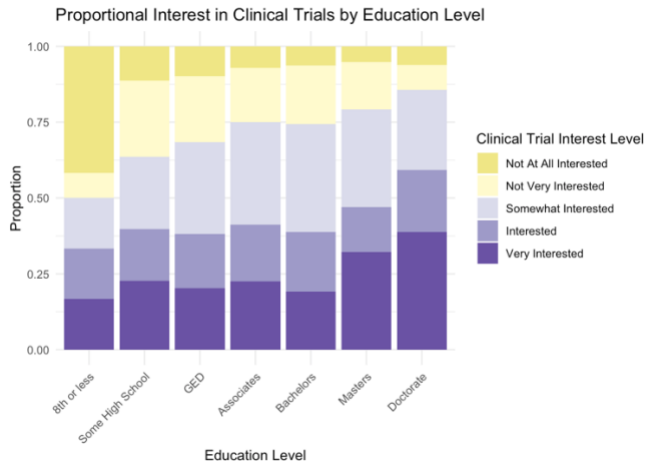
Preliminary Results

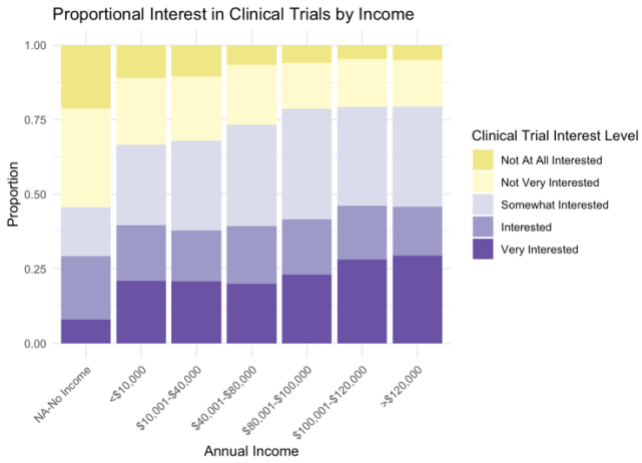


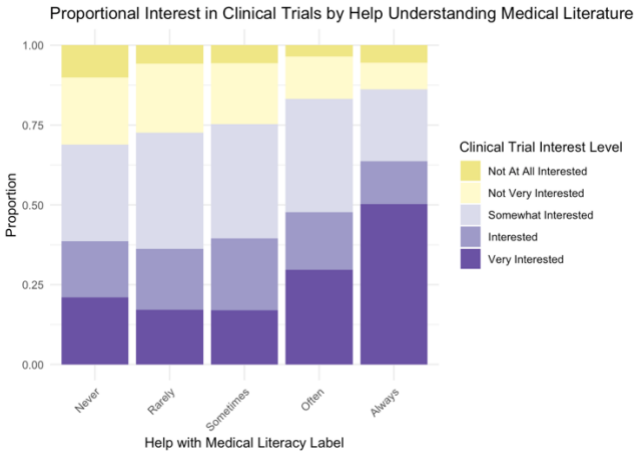












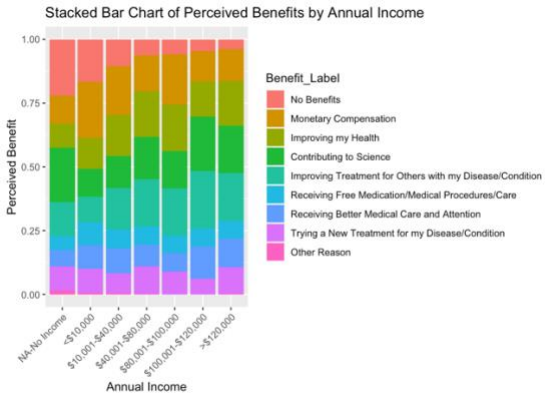
The preliminary results key findings: **Age and Gender**: The study's participants were diverse in terms of age and gender, with a wide age range and an almost equal distribution of males and females. **Geographic Variation**: Participants from this study come from 51 states in the US, relatively proportionally to each state's respective population. **Race**: The majority of participants in the study were White, which may affect the generalizability of the findings to more diverse populations. **Income Levels**: Most participants fell within the income range of $10,000 to $80,000, with income impacting interest in clinical trials. **Education**: Participants with higher education levels were more likely to express interest in clinical trials. **Health Insurance**: The majority of participants had health insurance, and interest in clinical trials was relatively consistent among those with and without insurance. **Years of Work Experience**: The distribution of work experience varied, with most participants having 10-40 years of experience.

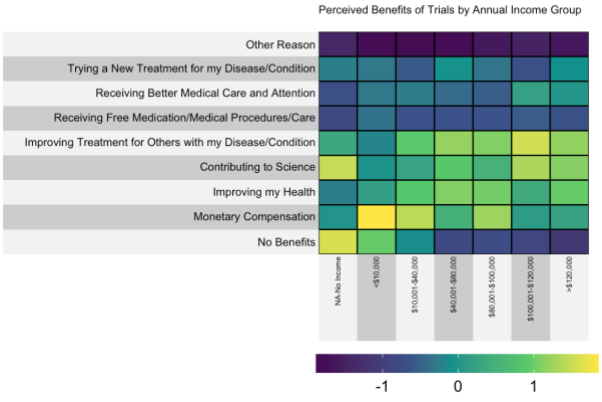
The main variable of interest, participant interest in clinical trials, showed that most participants were somewhat to very interested. The age group for this study ranges from 18-93 years old, with a mean age of 50. From the histogram and the quantile results we can see most people are between the ages 35-65 and it is mostly evenly distributed within that range. There is an even number of males and females (1990 males, 1995 females) with a small number of people identifying as another gender option (5 transgender male, 3 transgender female, 7 non-binary, 4 gender fluid and 2 other). We can see that we have a wide range of participants from each state. Unsurprisingly there are more participants from the larger states with a larger population (California, Florida, New York, Texas). This makes sense as these are the states in the US with the largest populations. See (https://www.statsamerica.org/sip/rank_list.aspx?rank_label=pop1). The majority of the participants in this study are White (3,089 out of 4,006). This is not representative of the entire country and will affect the analysis. Caution should be taken in interpreting this data in relation to other populations with a different distribution of race. Most people in this study are in the 10,000-40,000 or 40,001-80,000 Annual Income range (1,205 and 1,257 participants respectively). The average number of people in the household is ~2.5 and average number of dependents is ~1. The poverty threshold for the USA from 2022 is 14,880 for individuals, 18,990 for a household of 2, 23280 for a household of 3, 29950 for a household of 4, 35510

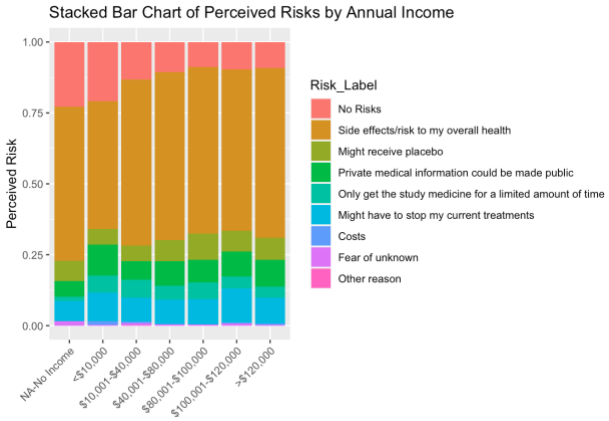
household of 5, 40160 for household of 6. Because we only have income ranges, we are not able to assess exact poverty levels relative to size of household. A majority of participants have a GED or bachelor's degree (1750 and 958 participants respectively).Most people in this study have health insurance (3561) Years of work experience ranges greatly from 0 to 74 years. From the histogram of work experience we can see that there are a large number of people with no work experience. The quantiles show that most people in the study population have worked for 10-40 years. Most people in this study either work full-time (1632 participants) or are retired (1116 participants).

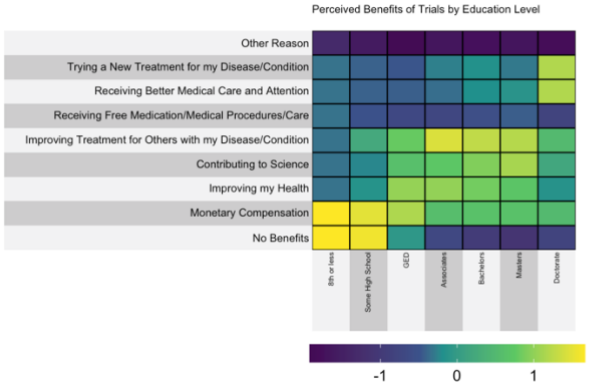
The main variable of interest in this analysis is participant's interest in participating in a clinical trial. Most people in the study replied they were 'somewhat interested' in participating in a clinical trial (1280 out of 4006). 329 replied 'not at all interest', 798 'not very interested', 728 'interested', and 871 'very interested'. So, most people in the study are somewhat to very interested.

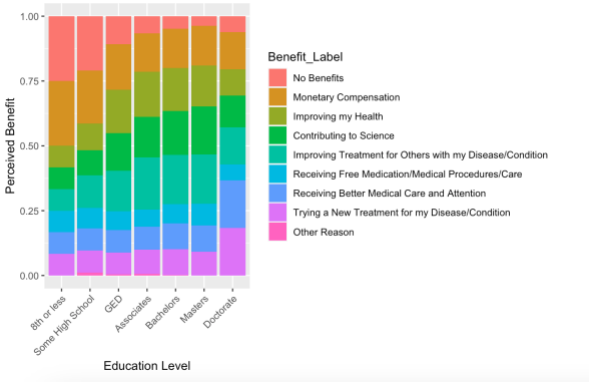
Interest levels are consistent compared to males and females. However, those participants that identified as Transgender Male/Female or Gender fluid were proportionally much more interested in participating in research. Interest levels are relatively consistent across employment status, with slightly more full-time employed people being somewhat to very interested in participating in a clinical trial. Interest levels are proportionally consistent across the different races and those who selected other. There is a slightly increase in participants interest levels as income increases, but this is very slight. However, those that reported no income had a proportionally significantly less interest in participating in clinical research. There is a steady increase in interest levels with increase in education levels with those participants with a doctorate having the highest levels of somewhat to very interested. Conversely, those who need additional help reading medical literature are more interested in participating in trials. English proficiency is a large factor in interest levels, proportionally those who have better English proficiency are more interested in research. Interestingly interest in clinical trials is consistent across living situations. Also consistent across location of living. The proportion of those interested in clinical trials is similar whether or not they have health insurance. Interest is also similar among people's ability to afford their medical care.

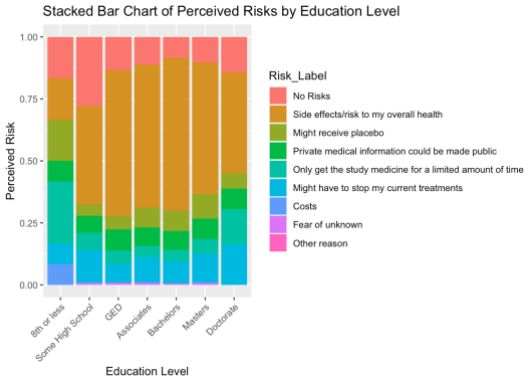


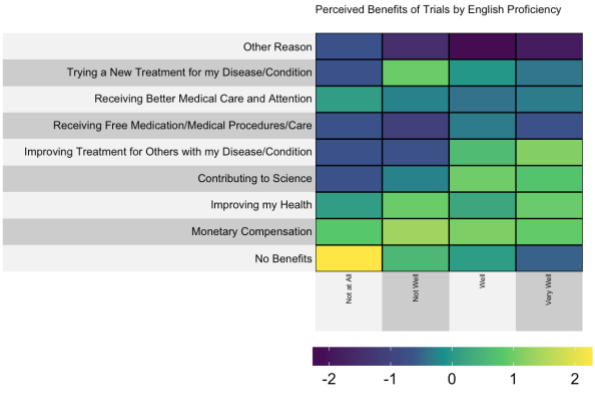


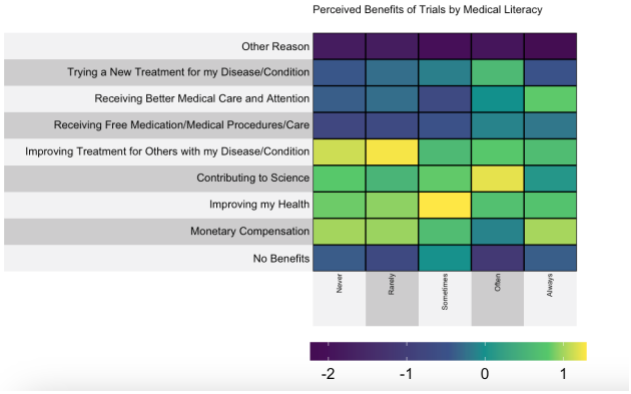

Stacked Bar Chart of Perceived Risks by Annual Income
Perceived Risk
1.00
0.75
0.50
0.25
0.00
NA-No Income
<$10,000
$10,001-$40,000
$40,001-$80,000
$80,001-$100,000
$100,001-$120,000
>$120,000
Risk_Label
No Risks
Side effects/risk to my overall health
Might receive placebo
Private medical information could be made public
Only get the study medicine for a limited amount of time
Might have to stop my current treatments
Costs
Fear of unknown
Other reason

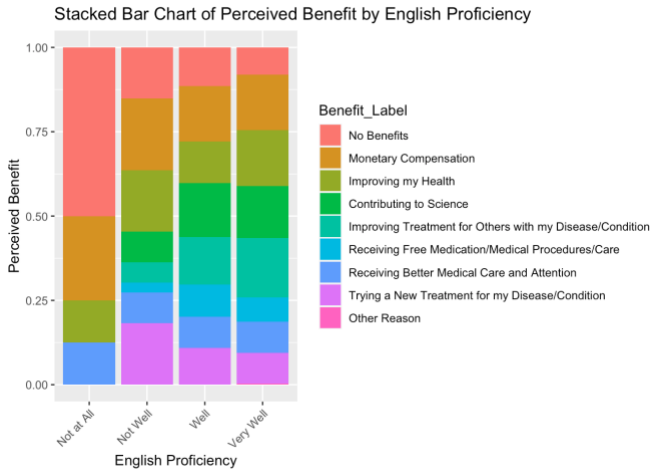

Perceived Benefits of Trials by Education Level
Other Reason
Trying a New Treatment for my Disease/Condition
Receiving Better Medical Care and Attention
Receiving Free Medication/Medical Procedures/Care
Improving Treatment for Others with my Disease/Condition
Contributing to Science
Improving my Health
Monetary Compensation
No Benefits
8th or less
Some High School
GED
Associates
Bachelors
Masters
Doctorate
-1
0
1

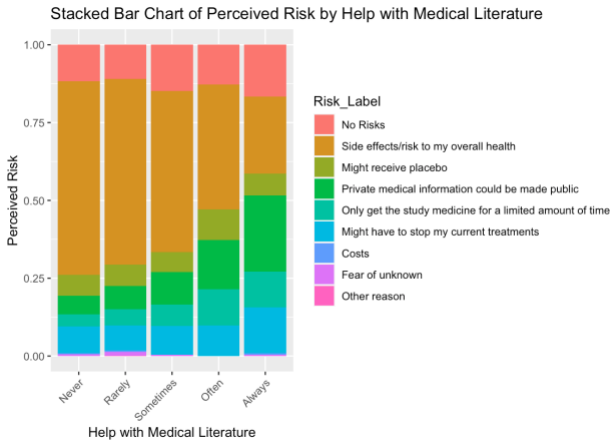

Stacked Bar Chart of Perceived Benefits by Education Level
Perceived Benefit
1.00
0.75
0.50
0.25
0.00
8th or less
Some High School
GED
Associates
Bachelors
Masters
Doctorate
Education Level
Benefit_Label
No Benefits
Monetary Compensation
Improving my Health
Contributing to Science
Improving Treatment for Others with my Disease/Condition
Receiving Free Medication/Medical Procedures/Care
Receiving Better Medical Care and Attention
Trying a New Treatment for my Disease/Condition
Other Reason


Stacked Bar Chart of Perceived Risks by Education Level
Perceived Risk
1.00
0.75
0.50
0.25
0.00
8th or less
Some High School
GED
Associates
Bachelors
Masters
Doctorate
Education Level
Risk_Label
No Risks
Side effects/risk to my overall health
Might receive placebo
Private medical information could be made public
Only get the study medicine for a limited amount of time
Might have to stop my current treatments
Costs
Fear of unknown
Other reason


Perceived Benefits of Trials by English Proficiency
Other Reason
Trying a New Treatment for my Disease/Condition
Receiving Better Medical Care and Attention
Receiving Free Medication/Medical Procedures/Care
Improving Treatment for Others with my Disease/Condition
Contributing to Science
Improving my Health
Monetary Compensation
No Benefits
Not at All
Not Well
Well
Very Well
-2
-1
0
1
2

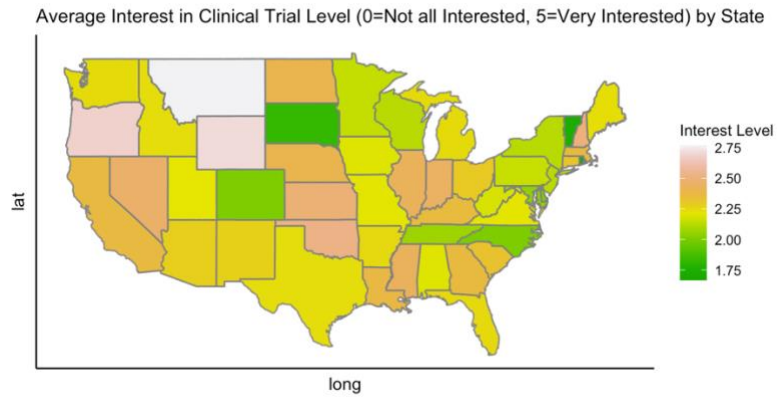

Perceived Benefits of Trials by Medical Literacy
Other Reason
Trying a New Treatment for my Disease/Condition
Receiving Better Medical Care and Attention
Receiving Free Medication/Medical Procedures/Care
Improving Treatment for Others with my Disease/Condition
Contributing to Science
Improving my Health
Monetary Compensation
No Benefits
Never
Rarely
Sometimes
Often
Always
-2
-1
0
1

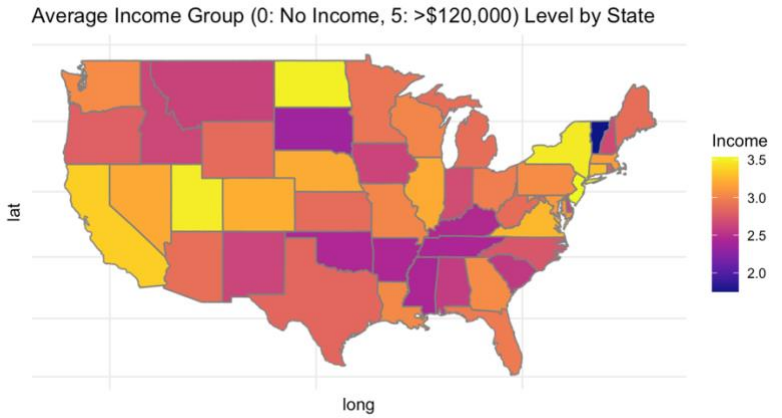




From these heatmaps we can see that for those in the lowest income group, <10,000, monetary compensation is the largest benefit for participating in a clinical study. The lower the income group, the higher rates of no perceived benefit. While side effects are the largest perceived risk across all income groups. Conversely, the lowest income groups also have the highest rates of no perceived risks. Perceived benefits and risks are similar among Males and Females. However, those participants that identified as Transgender Female or Gender fluid had the highest benefit of improving treatment for others. We can also see there is a clear trend with those with lowest levels of education selecting that there are no benefits to clinical trials or that the biggest benefit is monetary compensation. The 8th grade or less group also had the highest rate of only getting the study medicine for a limited amount of time, suggesting possible concerns with the ability to afford medical care.

Looking at English Proficiency and perceived benefits we can see that those participants that do not speak English at all have the highest perception of clinical trials providing no benefits. Interestingly, many people in the same English proficiency group also said they perceive no risks. Interestingly in this map, there is more of a spread of perceived benefits across ability to read medical literature, with equal spread in each group. Those who often or always need help with medical literature had a larger range of risks other than just side effects. Perceived risks and benefits were consistent across locations.





Vermont and South Dakota also had the lowest average interest in participating in clinical trials, supporting the theory that income is a factor in clinical trial participation. However, Colorado was one of the lower interest States but higher income. None of the states with the highest interest had the highest average income. Suggesting that lower income is more strongly correlated with clinical trial interest rather than higher income. This is not what we expected, as we saw that overall higher income levels resulted in higher interest levels. It seems that this phenomenon varies by state.

From this map we can see that the states with the highest average interest levels are Montana, Wyoming, Washington. The states with the lowest average interest levels are Vermont, South Dakota, Rhode Island, North Carolina, and Colorado. Looking at participant's average income by State, the states with the highest average income are North Dakota, Colorado, New York, New Jersey. The states with the lowest are Vermont, South Dakota, Mississippi, Oklahoma, Tennessee. Perceived risk seems consistent across states, specifically looking at the states at the extremes of income and interest. with side effect/risk to overall health being the primary risk concern. There is no clear trend that state influences clinical trial interest as they vary by income, education level, and other variables.

Conclusion

Based on the data we can see that there is evidence that those who are socioeconomically disadvantaged are less inclined to participate in clinical trials. People with lower education levels were less likely to see any medical or social benefits from participating in research studies, citing either no benefit or monetary benefits. Those participants in the highest income groups and higher education levels were more likely to see social/medical benefits for participating in clinical studies. Regardless of socioeconomic status, the biggest concern was side effects or risk to overall health. This was true across income levels, education levels, State, Medical Literacy. There were some contradictions within the groups of people who do not speak English well, perceiving no benefits but also no risk but consistent interest level compared to those who speak English well, suggesting some confusion. There doesn't seem to be a large difference in interest by location (rural, city, suburbs). However, there is does seem to be a difference by State.

Interactive features were added to enable users to explore and interact with the data easily, such as zooming in/out, hovering for specific information, and isolating data categories. The interactive graphs highlight the conclusions above. Figure 2 from the interactive plots allows to look at specific interest levels across different variables at once. Doing this is it is clear to see those with the lowest levels of education, income, English proficiency, and comprehension of medical literature are the least likely to participate. The trend is true for the opposite end of the spectrum as well, those with the highest levels of education, income, English proficiency, and comprehension of medical literature are the most likely to participate. This highlights the importance of even going 1 step up in socioeconomic status (i.e. from bachelor's to master's or often well to very well English skills) greatly affects interest in participating in clinical research. Figure 3 similarly allows you to investigate the data in a new way. Those with the lowest education levels view no benefits or largely monetary benefits. We can compare this to perceived benefits and see that low education subjects also do no view any benefits. So, we can conclude that those with lower education levels are mostly concerned with costs while higher education groups are concerned with side effects but see fewer risks overall. This highlights that the perception of clinical trials is largely different across socioeconomic groups. The risks and benefits for those in higher education groups or with higher literacy are most likely not actually different, however they are perceived to be different. We can see this clearly from interacting with Figure 3. Figure 4 is an interactive map , same as above, but allows for a clearer view and labeling of what statistics correspond to which state. It is interesting that the association we saw with higher income and higher interest isn't true here. We can see there are lower interests' level on the east side of the US compared to the left, however the states vary. Figure 5 explores this observation further. There is no clear positive or negative association with average income and average interest by state. This plot makes it easier to visualize the numeric difference from state to state. We can see most states are actually in the middle of the plot, suggesting moderate income and moderate interest. Some outliers we can see are Rhode Island, which has average income compared to the other states, but the lowest average interest. Contrasting Rhode Island, Alaska has one of the lowest levels of income but one of the highest average interests.

Overall, the data suggests that socioeconomically disadvantaged individuals, particularly those with lower education levels, income, and English language skills are less inclined to participate in clinical trials. This is reflected in their perception of fewer benefits and higher perceived risks. However, it is important to acknowledge the significance of side effects and health risks as primary concerns, consistent across various socioeconomic groups. While location (rural, city, suburbs) does not appear to significantly influence interest in clinical trials, the state of residence does. Perception varies greatly across the various socioeconomic groups and is perhaps the most interesting finding from the analysis. More research into why perceptions are so different would be an interesting next step, in addition to environmental factors that cause the trends to change at a state level.