Data Analysis Project Report

Dao Chi Tuong, Can Ha An, Nguyen Canh Huy

April 10, 2025

Introduction

The dataset we are using was explored in the July 2023 article *Racial and ethnic disparities in reproductive medicine in the United States: a narrative review of contemporary high-quality evidence*[1], published in the American Journal of Obstetrics and Gynecology in January 2025. The original paper focuses on racial and ethnic inequities in ob/gyn. The dataset suggestion is credited to Kat Correia from Amherst College.
This dataset provides a rich and structured source of information on medical research, particularly in the context of healthcare disparities and gynecological health outcomes, both article-level and model-level. It allows exploration of how different factors contribute to various gynecological conditions, quantifying disparities, and identifying key predictors of health outcomes.

Question 1: How are race and ethnicity categorized in medical research?

Introduction

This question examine the racial categories recorded and represented across studies, specifically those considered to be the primary groups in their respective works of research, as well as the studies that includes them. The specific parts of the dataset we use are the count occurrences to identify frequently reported racial/ethnic groups over time, plus the keywords extracted from the abstracts of the studies. This question is interesting because the underrepresentation of specific groups, as well as the imbalance in research topics between groups, should be studied.

Approach

To explore this question, we will use the following plots:

- **Plot 1:** A line chart for all race groups over the whole time period, as well as one bar chart for every individual year is used to visualize the number of articles that features the race group as the primary group. While the line chart can make the overall trends

and comparisons clear, the specific comparisons between groups with low numbers may not be effectively shown, and thus the bar chart would help specify this comparison.

- **Plot 2:** A word cloud is used to show the keywords of topics extracted from the articles featuring individual race groups. This allows us to visualize keywords in their respective importance of appearances.

Analysis

Below is the code used to generate the plots:

```
# Code Block 1
#Preprocessing columns
race_cols = [f"race{i}" for i in range(1, 9)]
ss_cols = [f"race{i}_ss" for i in range(1, 9)]
Q1_df=article_df.copy()
Q1_df.loc[:, ss_cols] = Q1_df[ss_cols].replace(-99, np.nan)

# Select only the desired columns
columns_to_keep = [
    "pmid", "doi", "jabbrv", "journal", "year", "month", "day",
    "title", "abstract", "keywords", "study_aim", "race1", "race1_ss
       "
]

Q1_df = Q1_df[columns_to_keep]

# Aggregate article counts per race_group and year
race_group_counts = Q1_df.groupby(["year", "race_group"]).size()
.reset_index(name="article_count")

min_year, max_year = int(df["year"].min()), int(df["year"].max())
# Get a fixed list of all race groups
all_race_groups = sorted(Q1_df["race_group"].unique())

# Initialize Dash app
app = dash.Dash(__name__)


app.layout = html.Div([
    dcc.Graph(id="race-group-line-chart", style={"width": "100%",
    "margin": "0 auto", "display": "block"}),

    html.Div([
        dcc.Slider(
            id="year-slider",
            min=min_year,
            max=max_year,
```

```
            value=max_year,  # Default to most recent year
            marks={year: str(year) for year in range(min_year,
            max_year + 1)},
            step=1
        ),
        html.Button("Reset", id="reset-button", n_clicks=0),
    ], id="slider-btn-div", style={"width": "50%", "margin":
    "0 auto", "textAlign": "center"}),

    html.Div([
        dcc.Graph(id="race-group-bar-chart", style={"width": "50%",
        "display": "inline-block"}),
        html.Img(id="word-cloud", style={"width": "50%", "display":
        "inline-block"}),
    ], style={"display": "flex", "justify-content": "center"}),
    html.Div()
])

@app.callback(
    [Output("race-group-bar-chart", "figure"),
     Output("race-group-line-chart", "figure"),
     Output("year-slider", "value")],
    [Input("year-slider", "value"),
     Input("reset-button", "n_clicks")]
)
def update_charts(selected_year, reset_clicks):
    triggered_id = ctx.triggered_id if ctx.triggered_id else "year-
       slider"

    if triggered_id == "reset-button" or not selected_year:
        filtered_df = race_group_counts.groupby("race_group")["
           article_count"]
        .sum().reset_index()
        bar_title = "Number of Articles by Race Group (All Years)"
        bar_chart = px.bar(filtered_df, x="race_group", y="
           article_count",
        title=bar_title)
        selected_year = max_year  # Reset slider
    else:
        filtered_df = race_group_counts[race_group_counts["year"] ==
            selected_year]
        full_data = pd.DataFrame({"race_group": all_race_groups})
        filtered_df = full_data.merge(filtered_df, on="race_group",
        how="left").fillna(0)
        filtered_df["article_count"] = filtered_df["article_count"]
        .astype(int)
```

```
        bar_title = f"Number of Articles by Race Group ({
            selected_year})"
        bar_chart = px.bar(filtered_df, x="race_group", y="
            article_count"
        , title=bar_title)

    line_chart = px.line(race_group_counts, x="year", y="
        article_count"
    , color="race_group", title="Articles Over Time by Race Group")
    line_chart.update_layout(legend=dict(font=dict(size=10)))  #
        Reduce legend font size

    return bar_chart, line_chart, selected_year

@app.callback(
    Output("word-cloud", "src"),
    Input("race-group-bar-chart", "clickData")
)
def update_word_cloud(click_data):
    if click_data is None:
        return None

    selected_race_group = click_data["points"][0]["x"]
    filtered_titles = Q1_df[Q1_df["race_group"] ==
        selected_race_group]["title"]
    text = " ".join(filtered_titles)

    wordcloud = WordCloud(width=400, height=300, background_color='
        white')
    .generate(text)
    img = BytesIO()
    wordcloud.to_image().save(img, format="PNG")
    encoded_img = base64.b64encode(img.getvalue()).decode()

    return f"{encoded_img}"


if __name__ == "__main__":
    app.run(port=8021,debug=True)
```

Figure 1: Line graph and Bar Chart for Race Groups and Articles

Figure 2: Word Cloud for Race Group by Year

As the plots are designed to be interactive, the full plots can not be shown in this report, and should be viewed as we run the code. A preview is included.

Discussion

First plot

- Overall Trends
  - White group consistently has the highest article count. Specifically, every year from 2010 to 2023, the "White" category consistently reports the most articles.

- Gradual increase in representation of other groups:
  - Early years (2010–2013) show minimal representation for groups other than White. By 2020 and beyond, more racial/ethnic groups begin appearing consistently in the data (e.g., Asian, Hispanic or Latino, American Indian or Alaska Native).
  - 2023 is notable for having the and the widest racial representation (6 different groups) with the most balanced numbers of articles.

Second plot

- Most of the topics are women-related, which is due to the data being collected from ob/gyn paper.
- "Racial" and "Disparities" is recorded in the keywords for most groups, which is understandable considering the data collected. However, such keywords are not shown for certain groups (e.g. Asians), meaning that while there are studies related to such topic, the topic itself have not been considered the point of focus to study on the group.
- "Cancer", being a complicated and thus focused health problem, is recorded in the keyword for all groups.
- Hispanic people is the only group with "Coronavirus" as a keyword. While that should correlate with the period of the studies (most of them from 2019-2023), more investigations can be made on this topic for this specific group.

Possible Interpretations

- Increased awareness and inclusion: Over time, there is a clear trend toward reporting on a more diverse set of racial/ethnic groups.
- Social or political influence: Peaks in certain years (e.g., 2020–2022) may correspond with real-world events, movements (e.g., BLM, COVID-19), or policy shifts that impacted media/reporting.
- More granular tracking: The appearance of new categories (like American Indian or Alaska Native) in later years could suggest improved or expanded data collection.

Question 2: What health outcomes have been studied, and what disparities have been identified?

Introduction

This question explores which health outcomes are most commonly studied in reproductive medicine and whether any disparities exist across racial or ethnic groups. Understanding which outcomes are prioritized can help identify gaps in research and uncover systematic biases. This analysis leverages a merged dataset containing article-level and model-level

metadata about studies on reproductive health. Specifically, we use health outcome classifications and reported effect sizes (e.g., odds ratios, risk ratios) to examine patterns across studies and demographic groups.

We're interested in this question because disparities in reproductive health outcomes — such as maternal mortality or neonatal ICU admissions — are acknowledged and documented in real-world clinical settings, but less is known about how thoroughly these disparities are studied and reported in the academic literature. This analysis provides an opportunity to explore that gap, and whether there is sufficient discussion regarding each racial group.

Approach

To explore this broad question, we break it down into 2 parts (each with its own visualizations):

2.1. Descriptive Analysis: What health outcomes have been studied?

- **Plot 1: Bubble chart** to visualize the frequency of different health outcome categories (e.g., maternal, neonatal, access to care) across all studies, as well as a **bar chart** to highlight most frequent health outcome per category, providing a more detailed understaning of most prevalent outcomes.
- **Plot 2: Heat map** to identify how racial groups are represented with respect to each health outcome.

2.2. Compare effect sizes and visualize disparities between different race groups

- **Plot 3: Violin Plot** to highlight the differences in the distribution and variability of effect sizes across racial and ethnic groups.
- **Plot 4: The Forest Plot**, with an interactive drop-down bar for race group selection, emphasizes outcome-specific disparities for each group.

Analysis

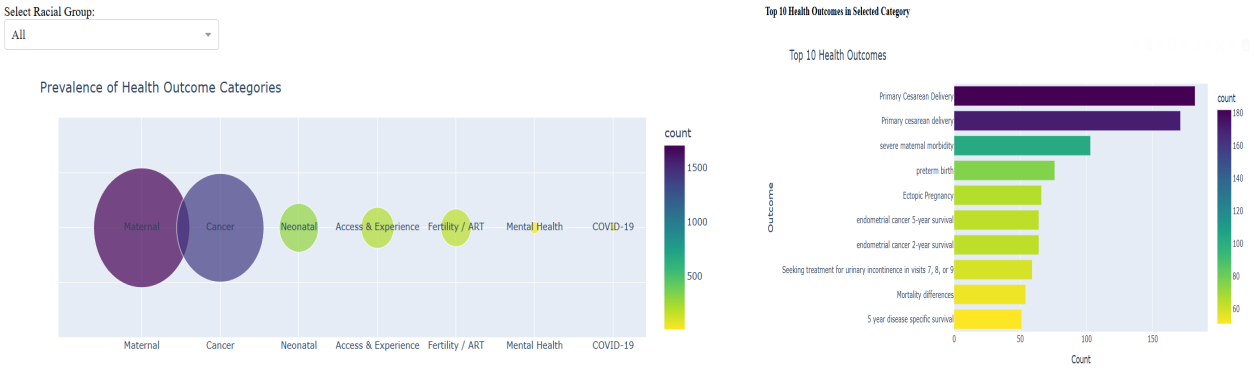
Code Block for Plot 1 + 2

```
from dash import Dash, html, dcc, Input, Output

df['race1_clean'] = standardize_race_labels(df['race1'])
app = Dash(__name__)

# --- Preprocess data for dropdown and plots ---
all_races = ['All'] + sorted(df['race1_clean'].dropna().unique().
   tolist())

category_counts_all = df['health_category'].value_counts().
   reset_index()
```

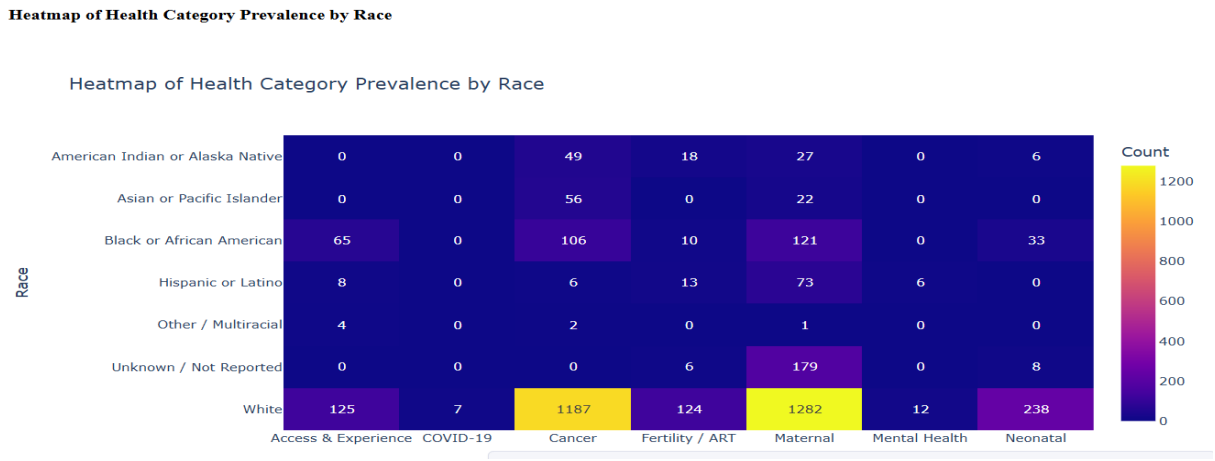
```
category_counts_all.columns = ['health_category', 'count']

# --- Layout ---
app.layout = html.Div([
    html.H2("Health Outcome Categories by Race"),

    html.Div([
        html.Label("Select Racial Group:"),
        dcc.Dropdown(
            id='race-selector',
            options=[{'label': race, 'value': race} for race in
               all_races],
            value='All',
            clearable=False
        )
    ], style={'width': '30%', 'marginBottom': '20px'}),

    dcc.Graph(id='category-bubble-chart'),

    html.H4("Top 10 Health Outcomes in Selected Category"),
    dcc.Graph(id='top-outcomes-bar'),

    html.H4("Heatmap of Health Category Prevalence by Race"),
    dcc.Graph(id='heatmap')
])

# --- Callbacks ---
@app.callback(
    Output('category-bubble-chart', 'figure'),
    Input('race-selector', 'value')
)
def update_bubble_chart(selected_race):
    if selected_race == 'All':
        data = df
    else:
        data = df[df['race1_clean'] == selected_race]

    category_counts = data[data['health_category'] != 'Other']['
       health_category'].value_counts().reset_index()
    category_counts.columns = ['health_category', 'count']

    fig = px.scatter(
        category_counts,
        x='health_category', y=[1]*len(category_counts),
        size='count', color='count', text='health_category',
        size_max=100, height=400, color_continuous_scale='Viridis_r'
    )
```

```
    fig.update_traces(textposition='middle center')
    fig.update_layout(
        showlegend=False,
        xaxis_title='', yaxis_title='',
        yaxis=dict(showticklabels=False),
        title="Prevalence of Health Outcome Categories"
    )
    return fig

@app.callback(
    Output('top-outcomes-bar', 'figure'),
    Input('race-selector', 'value')
)
def update_outcomes_bar(selected_race):
    if selected_race == 'All':
        data = df
    else:
        data = df[df['race1_clean'] == selected_race]

    outcome_counts = data['outcome'].value_counts().head(10).
       reset_index()
    outcome_counts.columns = ['outcome', 'count']
    outcome_counts = outcome_counts.sort_values('count', ascending=
       True)

    fig = px.bar(
        outcome_counts,
        x='count', y='outcome', orientation='h',
        color='count', color_continuous_scale='Viridis_r'
    )
    fig.update_layout(title="Top 10 Health Outcomes", yaxis_title="
       Outcome", xaxis_title="Count")
    return fig



@app.callback(
    Output('heatmap', 'figure'),
    Input('race-selector', 'value')
)
def update_heatmap(selected_race):
    data = df[df['health_category'] != 'Other']

    if selected_race != 'All':
        data = data[data['race1_clean'] == selected_race]
```

```
    heatmap_data = data.groupby(['race1_clean', 'health_category']).
        size().reset_index(name='count')
    pivot = heatmap_data.pivot(index='race1_clean', columns='
        health_category', values='count').fillna(0)

    fig = px.imshow(
        pivot,
        text_auto=True,
        labels=dict(x="Health Category", y="Race", color="Count"),
        title="Heatmap of Health Category Prevalence by Race",
        aspect="auto",
        height=500
    )
    return fig

# --- Run App ---
if __name__ == '__main__':
    app.run(port=8022, debug=True)
```



Plot 1: Bubble Chart and Bar Chart identifying Health Outcomes within Race Groups



Plot 2: Heat Map for Racial Group representation

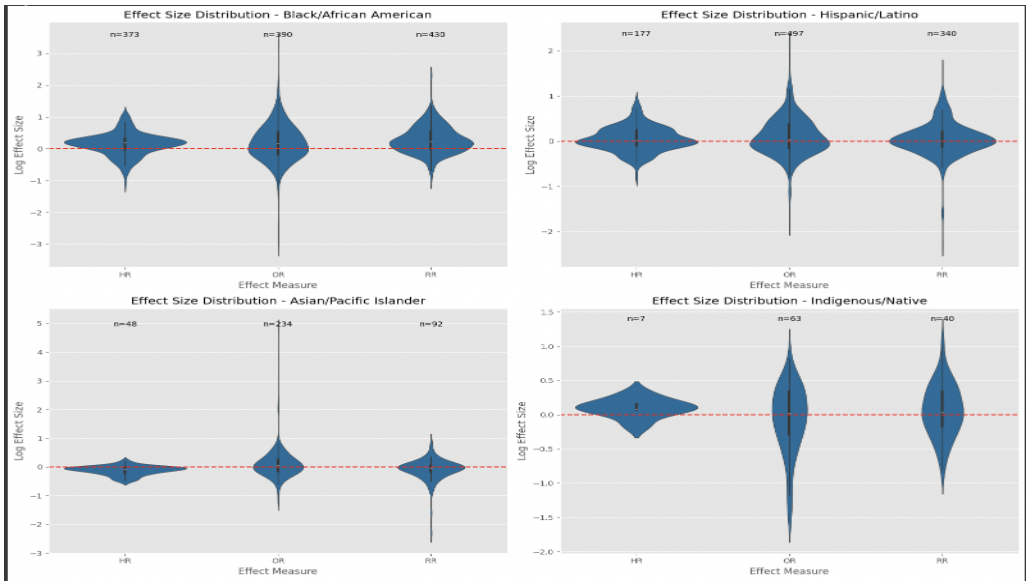
Code Block for Plot 3 + 4

```
# Initialize Dash app
app = dash.Dash(__name__)

# Define effect size measures
common_measures = ['OR', 'RR', 'HR']
effect_df = health_outcome_df[health_outcome_df['measure'].isin(
    common_measures)].copy()

# Log-transform effect sizes
effect_df['log_point'] = np.log(effect_df['point'].replace(0, np.nan
    ))

# Define demographic groups
demo_groups = {
    'is_black_comparison': 'Black/African American',
    'is_hispanic_comparison': 'Hispanic/Latino',
    'is_asian_comparison': 'Asian/Pacific Islander',
    'is_indigenous_comparison': 'Indigenous/Native'
}

# Dropdown options
race_options = [{"label": name, "value": key} for key, name in
    demo_groups.items()]

# Identify top outcomes
top_outcomes = effect_df['outcome'].value_counts().head(5).index.
    tolist()

# App layout with vertical stacking
app.layout = html.Div([
    html.H2("Effect Sizes by Demographic Group", style={"textAlign":
```

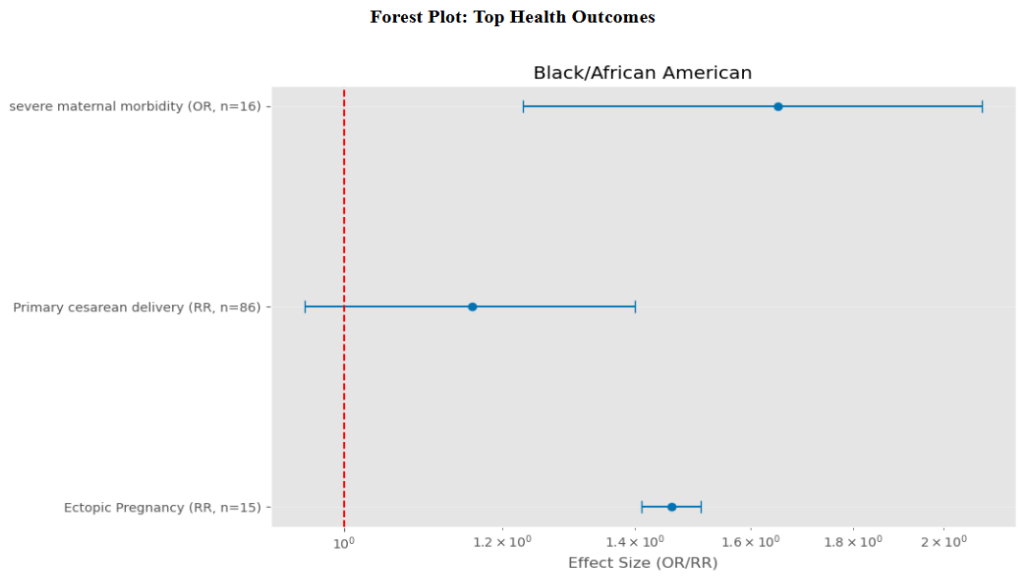
```
        "center"}),

    dcc.Dropdown(
        id="group-selector",
        options=race_options,
        value="is_black_comparison",
        style={"width": "50%", "margin": "0 auto 30px auto"}
    ),

    html.Div([
        html.H4("Violin Plot: Log Effect Size Distribution", style={
            "textAlign": "center"}),
        html.Img(id="violin-plot", style={"width": "80%", "margin":
            "auto", "display": "block", "marginBottom": "50px"}),
    ]),

    html.Div([
        html.H4("Forest Plot: Top Health Outcomes", style={"
            textAlign": "center"}),
        html.Img(id="forest-plot", style={"width": "80%", "margin":
            "auto", "display": "block"})
    ])
])

# Callback to update both plots
@app.callback(
    [Output("violin-plot", "src"),
     Output("forest-plot", "src")],
    Input("group-selector", "value")
)
def update_plots(group_key):
    group_name = demo_groups[group_key]
    group_data = effect_df[effect_df[group_key] == 1].copy()

    #### --- VIOLIN PLOT --- ####
    violin_src = None
    if not group_data.empty:
        plt.figure(figsize=(10, 8))
        ax = sns.violinplot(x='measure', y='log_point', data=
            group_data)
        ax.axhline(y=0, color='r', linestyle='--')
        ax.set_title(f'{group_name}')
        ax.set_ylabel('Log Effect Size')
        ax.set_xlabel('Effect Measure')
        for i, measure in enumerate(group_data['measure'].unique()):
            count = len(group_data[group_data['measure'] == measure
                ])
```

```
            ax.text(i, ax.get_ylim()[1]*0.9, f'n={count}', ha='
                center')
        buf_v = BytesIO()
        plt.tight_layout()
        plt.savefig(buf_v, format="png")
        plt.close()
        violin_src = f"{base64.b64encode(buf_v
            .getvalue()).decode()}"

#### --- FOREST PLOT --- ####
forest_src = None
forest_data = group_data[
    group_data['outcome'].isin(top_outcomes) &
    (~group_data['point'].isnull()) &
    (~group_data['lower'].isnull()) &
    (~group_data['upper'].isnull())
]

forest_stats = []
for outcome in top_outcomes:
    for measure in ['OR', 'RR']:
        subset = forest_data[
            (forest_data['outcome'] == outcome) &
            (forest_data['measure'] == measure)
        ]
        if len(subset) >= 3:
            forest_stats.append({
                "Outcome": outcome,
                "Measure": measure,
                "Count": len(subset),
                "Point": subset["point"].median(),
                "Lower": subset["lower"].median(),
                "Upper": subset["upper"].median()
            })

if forest_stats:
    forest_df = pd.DataFrame(forest_stats).sort_values(['Outcome
        ', 'Measure', 'Point'])
    plt.figure(figsize=(11, len(forest_df) + 3))
    y_pos = np.arange(len(forest_df))
    plt.errorbar(
        x=forest_df['Point'],
        y=y_pos,
        xerr=[forest_df['Point'] - forest_df['Lower'], forest_df
            ['Upper'] - forest_df['Point']],
        fmt='o', capsize=5
    )
```

```
        plt.yticks(y_pos, forest_df['Outcome'] + ' (' + forest_df['
            Measure'] + ', n=' + forest_df['Count'].astype(str) + ')'
            )
        plt.axvline(x=1, color='r', linestyle='--')
        plt.xscale('log')
        plt.xlabel('Effect Size (OR/RR)')
        plt.title(f'{group_name}')
        plt.grid(True, alpha=0.3)
        plt.tight_layout()
        buf_f = BytesIO()
        plt.savefig(buf_f, format="png")
        plt.close()
        forest_src = f"{base64.b64encode(buf_f
            .getvalue()).decode()}"

    return violin_src, forest_src

if __name__ == "__main__":
    app.run(debug=True, port=8023)
```



Plot 3: Violin Plot for Racial and Ethnic Groups



Plot 4: Forest Plot Showing Effect Size Measures

Discussion

1. The bar plot and bubble chart reveals that maternal outcomes are the most frequently studied health category, followed by neonatal outcomes and access to care. This indicates a strong research focus on mothers' health but potentially less emphasis on long-term outcomes or systemic barriers.

2. The heatmap reveals how different racial and ethnic groups are represented across various health outcome categories.

- **White participants** appear most frequently across nearly all health categories, confirming a strong over-representation of this group in the literature.
- **Black or African American** participants are also well represented (albeit much less than white participants), particularly in maternal outcomes, aligning with ongoing discussions around maternal health disparities.
- **Hispanic or Latino** and **Asian/Pacific Islander** populations are present in fewer categories, with much lower counts — indicating potential under-representation.
- **American Indian/Alaska Native** and **Multiracial or Other** groups are nearly absent in many categories, highlighting a critical research gap.

3. The Violin Plot highlights differences in the distribution and variability of effect sizes across racial and ethnic groups. Notably, studies involving Black or African American and American Indian/Alaska Native populations tend to report higher and more widely distributed log-transformed effect sizes, indicating both a greater magnitude of disparity and more variability across studies. In contrast, groups like White and Asian participants often show tighter distributions around smaller effect sizes, suggesting more consistent and

possibly more favorable outcomes. For better understanding the mechanism of the violin plot, we have:

- **Hazard Ratio (HR):** Measures the instantaneous risk of an event occurring in one group compared to another.
- **Odds Ratio (OR):** Compares the odds of an outcome occurring in one group versus another.
- **Risk Ratio (RR):** Directly compares the probability of an event occurring in an exposed group versus a non-exposed group.
- **The y-axis** shows the log effect size values. Values above 0 suggest increased risk/odds, values below 0 suggest decreased risk/odds, and values at 0 (red dotted line) indicate no effect.

The data shows consistent patterns of health disparities across demographic groups, with important nuances:

- **Black/African American population:**
  - The highest proportion of worse outcomes across all effect measures (OR: 59.0%, RR: 72.1%, HR: 72.9%)
  - Median effect sizes are consistently elevated (OR: 1.17, RR: 1.25, HR: 1.20)
  - The logistic regression model shows being in the Black comparison group is the only demographic predictor with a positive coefficient (0.369), meaning Black patients have 1.45 times higher odds of experiencing worse-than-median outcomes
- **Hispanic/Latino population:**
  - More moderate disparities with approximately half of studies showing worse outcomes
  - Median effect sizes close to null (OR: 1.04, RR: 1.01, HR: 1.01)
  - Negative coefficient in the predictive model (-0.378), suggesting lower odds of worse outcomes compared to reference
- **Asian/Pacific Islander population:**
  - Lower proportion of worse outcomes, especially for HR (18.8%)
  - Mean OR appears skewed (-99.00 to 120.10 range) but median shows slight disparities (1.04)
  - Strongest negative coefficient (-0.501) in the model, indicating significantly lower odds of worse outcomes
- **Indigenous/Native population:**
  - Mixed patterns with wide effect size ranges and limited sample sizes

- High proportion of HR studies showing worse outcomes (85.7%), but based on only 7 studies
- Near-zero coefficient in the model (-0.001), suggesting similar odds to reference after controlling for other factors

4. The Forest Plot adds a third layer of insight by visualizing individual effect sizes with confidence intervals, allowing for a direct comparison of disparities across studies. To be more specific, the forest plot for Black/African American population highlights outcome-specific disparities:

- **Ectopic Pregnancy**: RR of 1.46 [1.41, 1.51], representing a 46% increased risk
- **Primary cesarean delivery**: RR of 1.16 [0.95, 1.40], with confidence interval crossing null
- **Preterm birth**: OR of 1.46 [1.25, 2.01], showing significantly increased odds
- **Severe maternal morbidity**: OR of 1.65 [1.23, 2.09], the highest disparity magnitude

From the plot, it's evident that several studies involving minority racial groups report effect sizes significantly greater than 1, with relatively narrow confidence intervals — suggesting statistically significant disparities in outcomes like maternal morbidity, NICU admission, or access to treatment. Conversely, some studies show effect sizes close to or below 1, often involving White patients or more broadly defined populations.

Taken together, these visualizations suggest that while research has focused heavily on maternal and neonatal outcomes, it consistently reveals significant disparities affecting marginalized groups. The magnitude and variability of these disparities vary not only by race but also by the type of health outcome studied. These patterns highlight the importance of both broadening the scope of health outcomes examined and improving the consistency in how race and ethnicity are reported and analyzed in research.

Conclusion

This project explored racial and ethnic disparities in reproductive health research through two main questions: how race and ethnicity are categorized in medical studies, and what health outcomes have been studied in relation to racial disparities. Using a structured dataset derived from high-quality publications, we visualized trends in representation, keyword associations, and reported effect sizes.

Our first analysis revealed that studies continue to heavily focus on White populations, although representation of other groups—especially Black or African American—has increased over time. However, the range and focus of research topics still vary considerably by race group, with certain populations (e.g., Asian, Indigenous) underrepresented in both volume and topic specificity.

In the second analysis, we identified which health outcomes are most commonly studied, finding a dominant emphasis on maternal outcomes. The heatmap visualization highlighted stark differences in representation across racial groups, while violin and forest plots showed measurable disparities in effect sizes (particularly for marginalized populations) across key health indicators. These findings reflect both persistent inequities in health outcomes and limitations in how comprehensively these disparities are studied.

Overall, the analysis underscores the need for more inclusive and targeted research, especially for underrepresented racial and ethnic groups. Future work could integrate full-text data,—not just abstracts or keywords,—could offer deeper insights into framing and focus. Furthermore, we could evaluate intersectionality (e.g., race and socioeconomic status), or examine authorship patterns to explore structural bias in research production.

References

[1] Ayodele G. Lewis et al. "Racial and ethnic disparities in reproductive medicine in the United States: a narrative review of contemporary high-quality evidence". In: *American Journal of Obstetrics and Gynecology* 232.1 (Jan. 2025), 82–91.e44. ISSN: 0002-9378. DOI: 10.1016/j.ajog.2024.07.024. URL: http://dx.doi.org/10.1016/j.ajog.2024.07.024.