# Manuscript Tables and Figures

### Honoring self-designations on PA birth certificates

#### Josh Radack

make\_table <- **function**(ds, var, strat\_var){

tbl\_summary(

ds,

include = c(var),

by = strat\_var, *# split table by group*

percent = "row",

missing\_text = "Missing",

digits = list(all\_categorical() ~ c(0, 2))

) |>

add\_n() |> *# add column with total number of non-missing observations*

*# add\_p() |> # test for a difference between groups*

modify\_header(label = "\*\*Variable\*\*") |> *# update the column header*

bold\_labels()

}

make\_stratified\_table <- **function**(ds, var) {

ds %>%

select(one\_of(c(var, "hispynM", "ptb"))) %>%

tbl\_strata(

strata = hispynM,

.tbl\_fun =

~ .x %>%

tbl\_summary(

by = ptb,

percent = "row",

missing\_text = "Missing",

digits = list(all\_categorical() ~ c(0, 2))

) %>%

modify\_column\_hide(columns = "stat\_1") %>%

add\_overall(statistic = everything() ~ "{n}"),

.header = "\*\*{strata}\*\*"

) |>

modify\_footnote(c(stat\_0\_1, stat\_0\_2) ~ "n")

}

top\_indeterminate <- **function**(hispanic, count = 20) {

pa\_data |>

filter(hispynM == hispanic & race\_write\_in\_cleaned == "indeterminate") |>

count(mom\_other\_literal) |>

arrange(-n) |>

head(n=count)

}

substitute\_labels <- **function**(x) {

case\_match(

x,

"american indian/native alaskan" ~ "American Indian/Alaskan Native",

"asian" ~ "Asian",

"black" ~ "Black",

"multiple" ~ "Multiple",

"native hawaiian/pacific islander" ~ "Native Hawaiian/Pacific Islander",

"white" ~ "White",

"africa-other" ~ "Africa",

"europe-other" ~ "Europe",

"mena-other" ~ "Middle East/North Africa",

"multiple continents-other" ~ "Multiple Continents",

"north america-other" ~ "North America",

"oceania-other" ~ "Oceania",

"south america-other" ~ "South America",

"indeterminate" ~ "Indeterminate",

"other" ~ "Other",

"other (could not reassign)" ~ "Other (Could Not Reassign)",

"other (no write-in)" ~ "Other (No Write-In)",

"no write-in" ~ "No Write-In",

.default = x

)

}

race\_nice\_label <- c(

"American Indian/Alaskan Native" = "american indian/native alaskan",

"Asian" = "asian",

"Black" = "black",

"Multiple" = "multiple",

"Native Hawaiian/Pacific Islander" = "native hawaiian/pacific islander",

"White" = "white",

"Africa" = "africa-other",

"Europe" = "europe-other",

"Middle East/North Africa" = "mena-other",

"Multiple Continents" = "multiple continents-other",

"North America" = "north america-other",

"Oceania" = "oceania-other",

"South America" = "south america-other",

"Indeterminate" = "indeterminate"

)

Hide

fit\_poisson <- **function**(hispanic\_strata, ref\_level, filter\_out) {

pa\_data %>%

filter(!race\_write\_in\_cleaned %**in**% filter\_out) %>%

filter(hispynM == hispanic\_strata) %>%

mutate(

ptb\_ind = if\_else(ptb == "PTB", 1, 0),

race\_write\_in\_cleaned = race\_write\_in\_cleaned %>%

as.factor() %>% relevel(ref = ref\_level)

) %>%

glm(ptb\_ind ~ race\_write\_in\_cleaned, data = ., family = poisson())

}

mod\_summary <- **function**(mod, reference\_group) {

mod %>%

broom::tidy(conf.int = TRUE) %>%

mutate(

term = gsub("race\_write\_in\_cleaned" , "", term),

RR = exp(estimate),

conf.low = exp(conf.low),

conf.high = exp(conf.high)

) %>%

select(

term, RR, std.error, statistic, p.value, conf.low, conf.high

) %>%

add\_row(

term = reference\_group,

RR = 1,

.before = 1

)

}

mod\_summary\_table <- **function**(mod\_sum) {

mod\_sum %>%

kableExtra::kable() %>%

kableExtra::kable\_styling()

}

mod\_forest\_plot <- **function**(mod\_sum) {

mod\_sum %>%

filter(term != "(Intercept)") %>%

mutate(

term = factor(term, levels = race\_factor\_levels) %>%

fct\_recode(

!!!race\_nice\_label

)

) %>%

ggplot(aes(x=fct\_rev(term), y=RR, ymin=conf.low, ymax=conf.high)) +

geom\_pointrange() +

geom\_hline(yintercept=1, lty=2) + *# add a dotted line at x=1 after flip*

coord\_flip() + *# flip coordinates (puts labels on y axis)*

scale\_y\_continuous(trans = "log", labels = \(x) sprintf("%.2f", x)) + *# converts to logarithmic scale*

xlab("Racial and Continental Origin Identity") + ylab("RR (95% CI)") +

theme\_bw() + *# use a white background*

theme(axis.text.y = element\_text(size = 10))

}

# Data Load and Clean

Begin by loading and cleaning the original PA birth ceritificates dataset and the crosswalk for write-in responses.

Hide

race\_factor\_levels <- c(

"american indian/native alaskan",

"asian",

"black",

"indigenous",

"multiple",

"native hawaiian/pacific islander",

"white",

"africa-other",

"europe-other",

"mena-other",

"multiple continents-other",

"north america-other",

"oceania-other",

"south america-other",

"indeterminate"

)

pa\_data <- readRDS("data/pa\_data\_merged\_write\_in\_20240223.rds") |>

mutate(

race\_write\_in\_cleaned = race\_write\_in\_cleaned |>

factor(levels = race\_factor\_levels)

)

var\_label(pa\_data) <- list(

racem\_me\_mapped = "Maternal Race Mapped",

race\_write\_in\_cleaned = "Maternal Race Write-In Cleaned",

racem\_me\_write\_in = "Maternal Race w/Write-In",

racem\_me\_write\_in\_collapsed = "Maternal Race w/Write-In (Collapsed)",

mage\_cat = "Maternal Age",

nulliparous = "Nulliparous",

maternal\_education = "Maternal Education",

smoke\_any = "Any Smoking",

bc\_mombmipre = "BMI (Prepregnancy)",

hispynM = "Hispanic (Y/N)",

married = "Married"

)

# Comparison of Single and Multiple Choice Race Variables

The following tables compare the number of respondents who selected “Other” in the single choice and multiple choice race variables. We see that fewer people selected “Other” when presented with multiple choices.

Hide

pa\_data %>%

mutate(

single\_choice\_race = if\_else(mom\_single\_race == "Other", "Other", "Not Other")

) %>%

count(single\_choice\_race)

## single\_choice\_race n

## 1 Not Other 1117857

## 2 Other 78268

Hide

pa\_data %>%

mutate(

multiple\_choice\_race = if\_else(racem\_me\_mapped == "other", "Other", "Not Other")

) %>%

count(multiple\_choice\_race)

## multiple\_choice\_race n

## 1 Not Other 1123234

## 2 Other 72891

The following table shows the original options for the multiple choice race variable.

Hide

pa\_data %>%

count(racem\_me\_mapped)

## racem\_me\_mapped n

## 1 american indian/native alaskan 1017

## 2 asian 46165

## 3 black 171965

## 4 multiple 40802

## 5 native hawaiian/pacific islander 726

## 6 other 72891

## 7 unknown 20901

## 8 white 841658

# Count of the number of who click “other” and whether they had a write-in or not

This table shows the number of people who had a write-in and didn’t have a write-in response among those who selected “other” as their race.

Hide

pa\_data |>

filter(racem\_me\_mapped == "other") |>

count(other\_write\_in)

## other\_write\_in n

## 1 has write-in 72744

## 2 no write-in 147

# Tables

## Table 1. Cohort summary of race and ethnicity

This table shows the maternal race based on what was checked on the birth certificate. The multiple choice/check box version of the race variable was used, with entries binned together for Asian and Native Hawaiian/Pacific Islander.

Hide

pa\_data %>%

select(one\_of(c("racem\_me\_mapped", "hispynM"))) %>%

tbl\_summary(

by = hispynM,

percent = "column",

digits = list(all\_categorical() ~ c(0, 2))

)

| **Characteristic** | **Hispanic**, N = 119,337*1* | **Non-Hispanic**, N = 1,076,788*1* |
| --- | --- | --- |
| Maternal Race Mapped |  |  |
| american indian/native alaskan | 400 (0.34%) | 617 (0.06%) |
| asian | 584 (0.49%) | 45,581 (4.23%) |
| black | 4,120 (3.45%) | 167,845 (15.59%) |
| multiple | 8,215 (6.88%) | 32,587 (3.03%) |
| native hawaiian/pacific islander | 350 (0.29%) | 376 (0.03%) |
| other | 65,073 (54.53%) | 7,818 (0.73%) |
| unknown | 16,925 (14.18%) | 3,976 (0.37%) |
| white | 23,670 (19.83%) | 817,988 (75.97%) |
| *1* n (%) | | |

## Table 2. Preterm birth rates among individuals who chose “other” to the race question and added write-in responses

This table recreates table 1 but only among those who identified as “Other”. Those who did not identify as “Other” are not included. The counts of “Missing” correspond to those who did not have a write-in response.

Hide

pa\_data %>%

filter(racem\_me\_mapped == "other") %>%

make\_stratified\_table("race\_write\_in\_cleaned") |>

as\_gt() |>

gt::tab\_style(style = cell\_borders(sides = c("bottom"), weight = px(0.5)),

locations = cells\_body(rows = c(8, 15)))

| **Characteristic** | **Hispanic** | | **Non-Hispanic** | |
| --- | --- | --- | --- | --- |
| **Overall**, N = 65,073*1* | **PTB**, N = 5,669*2* | **Overall**, N = 7,818*1* | **PTB**, N = 587*2* |
| Maternal Race Write-In Cleaned |  |  |  |  |
| american indian/native alaskan | 28 | 3 (10.71%) | 23 | 4 (17.39%) |
| asian | 105 | 11 (10.48%) | 2,028 | 139 (6.85%) |
| black | 90 | 9 (10.00%) | 115 | 7 (6.09%) |
| indigenous | 0 | 0 (NA%) | 0 | 0 (NA%) |
| multiple | 193 | 11 (5.70%) | 369 | 39 (10.57%) |
| native hawaiian/pacific islander | 3 | 0 (0.00%) | 4 | 1 (25.00%) |
| white | 126 | 12 (9.52%) | 62 | 1 (1.61%) |
| africa-other | 27 | 3 (11.11%) | 602 | 43 (7.14%) |
| europe-other | 252 | 26 (10.32%) | 480 | 29 (6.04%) |
| mena-other | 160 | 10 (6.25%) | 2,166 | 133 (6.14%) |
| multiple continents-other | 172 | 8 (4.65%) | 202 | 17 (8.42%) |
| north america-other | 18,838 | 1,736 (9.22%) | 889 | 91 (10.24%) |
| oceania-other | 0 | 0 (NA%) | 1 | 0 (0.00%) |
| south america-other | 1,467 | 90 (6.13%) | 242 | 30 (12.40%) |
| indeterminate | 43,495 | 3,735 (8.59%) | 605 | 52 (8.60%) |
| Missing | 117 | 15 | 30 | 1 |
| *1* n | | | | |
| *2* n (%) | | | | |

Here is a tabulation of original write-in entry for those classified as “indigenous.”

Hide

pa\_data |>

filter(race\_write\_in\_cleaned == "indigenous") |>

count(mom\_other\_literal\_orig)

## [1] mom\_other\_literal\_orig n

## <0 rows> (or 0-length row.names)

Hide

pa\_data |>

filter(race\_write\_in\_cleaned == "asia-other") |>

select(mom\_other\_literal, KH\_Race, KH\_Continent\_unique)

## Supplemental table 1 - Preterm birth rates by race and ethnicity

The following tables show the preterm birth rates by race and ethnicity. The first table is before any of the reassignments occurred, the latter once individuals who chose “other” and provided a written response were added to preexisting bins.

### Pre-Reassigment

Hide

pa\_data %>%

make\_stratified\_table("racem\_me\_mapped")

| **Characteristic** | **Hispanic** | | **Non-Hispanic** | |
| --- | --- | --- | --- | --- |
| **Overall**, N = 119,337*1* | **PTB**, N = 10,207*2* | **Overall**, N = 1,076,788*1* | **PTB**, N = 84,513*2* |
| Maternal Race Mapped |  |  |  |  |
| american indian/native alaskan | 400 | 48 (12.00%) | 617 | 57 (9.24%) |
| asian | 584 | 48 (8.22%) | 45,581 | 2,882 (6.32%) |
| black | 4,120 | 415 (10.07%) | 167,845 | 18,988 (11.31%) |
| multiple | 8,215 | 796 (9.69%) | 32,587 | 2,885 (8.85%) |
| native hawaiian/pacific islander | 350 | 34 (9.71%) | 376 | 31 (8.24%) |
| other | 65,073 | 5,669 (8.71%) | 7,818 | 587 (7.51%) |
| unknown | 16,925 | 1,454 (8.59%) | 3,976 | 398 (10.01%) |
| white | 23,670 | 1,743 (7.36%) | 817,988 | 58,685 (7.17%) |
| *1* n | | | | |
| *2* n (%) | | | | |

### Post-Reassignment

Hide

pa\_data %>%

mutate(

racem\_me\_write\_in\_collapsed = if\_else(grepl("other", racem\_me\_write\_in\_collapsed, fixed = TRUE), "other", racem\_me\_write\_in\_collapsed)

) %>%

make\_stratified\_table("racem\_me\_write\_in\_collapsed")

| **Characteristic** | **Hispanic** | | **Non-Hispanic** | |
| --- | --- | --- | --- | --- |
| **Overall**, N = 119,337*1* | **PTB**, N = 10,207*2* | **Overall**, N = 1,076,788*1* | **PTB**, N = 84,513*2* |
| racem\_me\_write\_in\_collapsed |  |  |  |  |
| american indian/native alaskan | 428 | 51 (11.92%) | 640 | 61 (9.53%) |
| asian | 689 | 59 (8.56%) | 47,609 | 3,021 (6.35%) |
| black | 4,210 | 424 (10.07%) | 167,960 | 18,995 (11.31%) |
| multiple | 8,408 | 807 (9.60%) | 32,956 | 2,924 (8.87%) |
| native hawaiian/pacific islander | 350 | 34 (9.71%) | 376 | 31 (8.24%) |
| other | 64,531 | 5,623 (8.71%) | 5,221 | 397 (7.60%) |
| unknown | 16,925 | 1,454 (8.59%) | 3,976 | 398 (10.01%) |
| white | 23,796 | 1,755 (7.38%) | 818,050 | 58,686 (7.17%) |
| *1* n | | | | |
| *2* n (%) | | | | |

### Supplemental table 1 PTB rate comparisons

The following table compares the preterm birth rate before and after write-ins are reclassified.

Hide

ptb\_rate\_tab <- **function**(var\_name, res\_name) {

pa\_data |>

group\_by({{ var\_name }}, hispynM) |>

summarize(

"{{ res\_name }}" := sum(if\_else(ptb == "PTB", 1, 0)) / n()

)

}

table1\_ptb\_rates <- ptb\_rate\_tab(racem\_me\_mapped, ptb\_rate\_pre)

## `summarise()` has grouped output by 'racem\_me\_mapped'. You can override using

## the `.groups` argument.

Hide

pa\_data <- pa\_data %>%

mutate(

racem\_me\_write\_in\_collapsed\_v2 = if\_else(racem\_me\_write\_in\_collapsed %**in**% c("other (could not reassign)", "other (no write-in)"), "other (could not reassign or no write-in)", racem\_me\_write\_in\_collapsed)

)

table3\_ptb\_rates <- ptb\_rate\_tab(racem\_me\_write\_in\_collapsed\_v2, ptb\_rate\_post)

## `summarise()` has grouped output by 'racem\_me\_write\_in\_collapsed\_v2'. You can

## override using the `.groups` argument.

Hide

tab\_comp <- table1\_ptb\_rates |>

full\_join(

table3\_ptb\_rates,

by = c("racem\_me\_mapped" = "racem\_me\_write\_in\_collapsed\_v2", "hispynM" = "hispynM")

) |>

mutate(ptb\_rate\_diff = ptb\_rate\_post - ptb\_rate\_pre) |>

mutate(

across(

c(ptb\_rate\_pre, ptb\_rate\_post, ptb\_rate\_diff),

~ if\_else(is.na(.x), NA, sprintf("%1.2f%%", 100 \* .x))

)

) |>

rename(race = racem\_me\_mapped)

tab\_comp |>

mod\_summary\_table()

| **race** | **hispynM** | **ptb\_rate\_pre** | **ptb\_rate\_post** | **ptb\_rate\_diff** |
| --- | --- | --- | --- | --- |
| american indian/native alaskan | Hispanic | 12.00% | 11.92% | -0.08% |
| american indian/native alaskan | Non-Hispanic | 9.24% | 9.53% | 0.29% |
| asian | Hispanic | 8.22% | 8.56% | 0.34% |
| asian | Non-Hispanic | 6.32% | 6.35% | 0.02% |
| black | Hispanic | 10.07% | 10.07% | -0.00% |
| black | Non-Hispanic | 11.31% | 11.31% | -0.00% |
| multiple | Hispanic | 9.69% | 9.60% | -0.09% |
| multiple | Non-Hispanic | 8.85% | 8.87% | 0.02% |
| native hawaiian/pacific islander | Hispanic | 9.71% | 9.71% | 0.00% |
| native hawaiian/pacific islander | Non-Hispanic | 8.24% | 8.24% | 0.00% |
| other | Hispanic | 8.71% | NA | NA |
| other | Non-Hispanic | 7.51% | NA | NA |
| unknown | Hispanic | 8.59% | 8.59% | 0.00% |
| unknown | Non-Hispanic | 10.01% | 10.01% | 0.00% |
| white | Hispanic | 7.36% | 7.38% | 0.01% |
| white | Non-Hispanic | 7.17% | 7.17% | -0.00% |
| other (could not reassign or no write-in) | Hispanic | NA | 8.71% | NA |
| other (could not reassign or no write-in) | Non-Hispanic | NA | 7.60% | NA |

# Association of write-in with preterm birth

We test for whether there is a statistically significant difference in the preterm birth rate between the different write-in groups.

## Poisson Rate Ratio Models

For the Hispanic patients, we use “north america-other” as the reference level. Those who did not originally choose “Other” and those who were classified as “indeterminate,” “oceania-other,” or “native hawaiian/pacific islander” were excluded (the latter two due to small sample size).

Hide

hispanic\_poisson <- fit\_poisson("Hispanic", "north america-other", c("indeterminate", NA, "oceania-other", "native hawaiian/pacific islander"))

hispanic\_poisson\_mod\_sum <- hispanic\_poisson %>%

mod\_summary("north america-other")

hispanic\_poisson\_mod\_sum %>%

mod\_summary\_table()

| **term** | **RR** | **std.error** | **statistic** | **p.value** | **conf.low** | **conf.high** |
| --- | --- | --- | --- | --- | --- | --- |
| north america-other | 1.0000000 | NA | NA | NA | NA | NA |
| (Intercept) | 0.0927425 | 0.0223328 | -106.4770493 | 0.0000000 | 0.0887420 | 0.0968614 |
| american indian/native alaskan | 1.0782544 | 0.4088587 | 0.1842774 | 0.8537959 | 0.4281444 | 2.1880958 |
| asian | 0.9514009 | 0.2895377 | -0.1720665 | 0.8633852 | 0.5087279 | 1.5987451 |
| black | 1.1275863 | 0.2509955 | 0.4784121 | 0.6323569 | 0.6602625 | 1.7777307 |
| multiple | 0.7449757 | 0.2305002 | -1.2772383 | 0.2015181 | 0.4573558 | 1.1346622 |
| white | 1.2611162 | 0.2247193 | 1.0323868 | 0.3018909 | 0.7845292 | 1.9020299 |
| africa-other | 0.9802312 | 0.5777820 | -0.0345576 | 0.9724325 | 0.2436105 | 2.5439433 |
| europe-other | 1.2878448 | 0.1417977 | 1.7840210 | 0.0744203 | 0.9629166 | 1.6807014 |
| mena-other | 0.7300681 | 0.2782478 | -1.1307100 | 0.2581772 | 0.4010056 | 1.2042889 |
| multiple continents-other | 0.6392812 | 0.2591629 | -1.7263691 | 0.0842810 | 0.3672730 | 1.0217181 |
| south america-other | 0.6899824 | 0.0979268 | -3.7894540 | 0.0001510 | 0.5662493 | 0.8315355 |

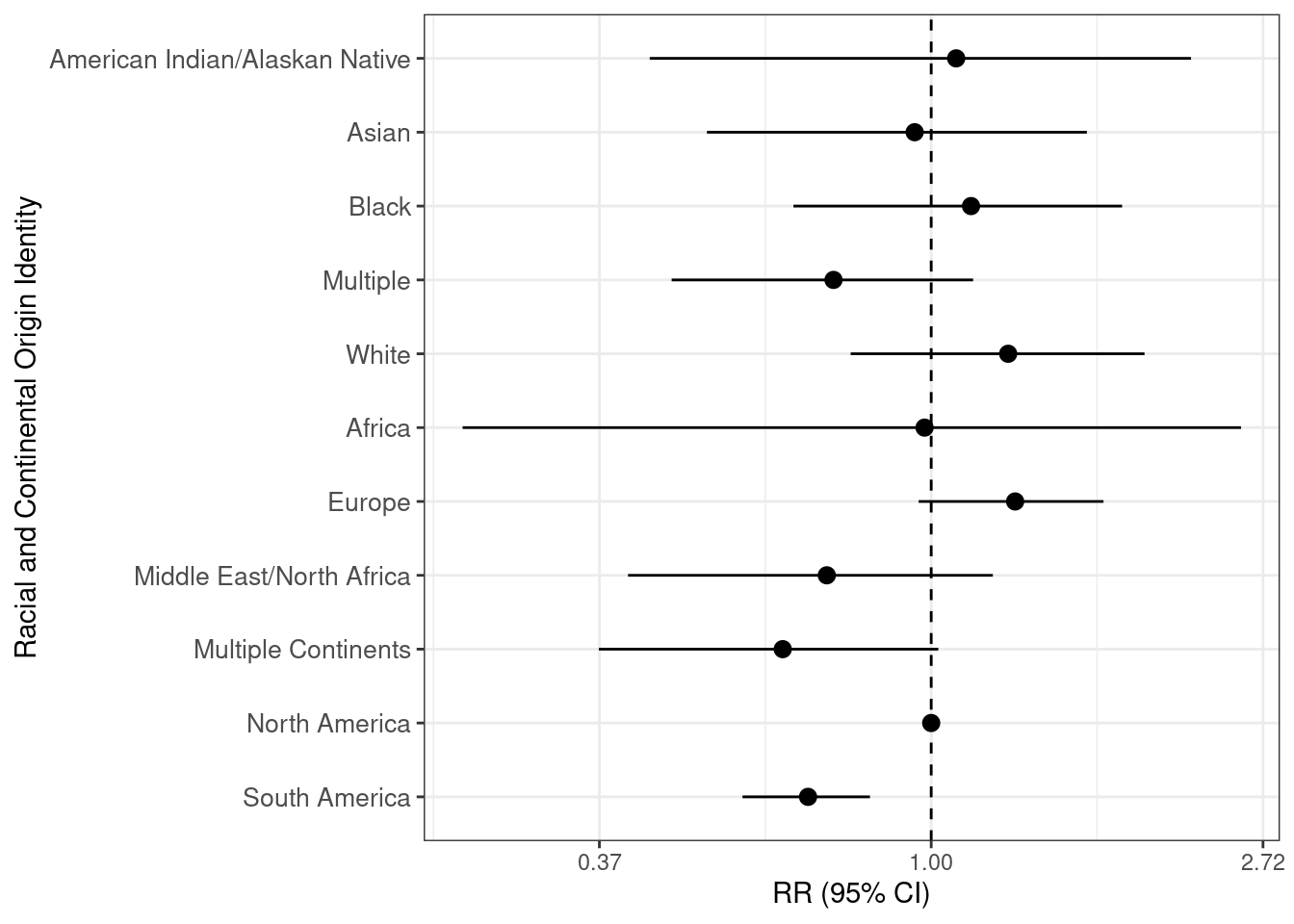
Hide

hispanic\_poisson\_fp <- hispanic\_poisson\_mod\_sum %>%

mod\_forest\_plot()

hispanic\_poisson\_fp

## Warning: Removed 1 rows containing missing values (`geom\_segment()`).



Hide

ggsave("tables\_figures/hispanic\_poisson\_forest\_plot.jpeg", hispanic\_poisson\_fp, width = 8, height = 6, dpi = 300, units = "in")

## Warning: Removed 1 rows containing missing values (`geom\_segment()`).

For the non-Hispanic patients, we use “asian” as the reference level. Those who did not originally choose “Other” and those who were classified as “indeterminate,” “indigenous,” “oceania-other,” or “native hawaiian/pacific islander” were excluded (the latter three due to small smaple sizes).

Hide

non\_hispanic\_poisson <- fit\_poisson("Non-Hispanic", "asian", c("indeterminate", NA, "indigenous", "oceania-other", "native hawaiian/pacific islander"))

non\_hispanic\_poisson\_mod\_sum <- non\_hispanic\_poisson %>%

mod\_summary("asian")

non\_hispanic\_poisson\_mod\_sum %>%

mod\_summary\_table()

| **term** | **RR** | **std.error** | **statistic** | **p.value** | **conf.low** | **conf.high** |
| --- | --- | --- | --- | --- | --- | --- |
| asian | 1.0000000 | NA | NA | NA | NA | NA |
| (Intercept) | 0.0761956 | 0.0595491 | -43.2323921 | 0.0000000 | 0.0676449 | 0.0854377 |
| american indian/native alaskan | 2.0265175 | 0.2261971 | 3.1225807 | 0.0017927 | 1.2615425 | 3.0757730 |
| black | 1.1523612 | 0.2431083 | 0.5833328 | 0.5596693 | 0.6900246 | 1.7997148 |
| multiple | 1.2054139 | 0.1215815 | 1.5366078 | 0.1243894 | 0.9448609 | 1.5227066 |
| white | 0.4059004 | 0.5804125 | -1.5534598 | 0.1203133 | 0.1005340 | 1.0611259 |
| africa-other | 0.9547401 | 0.1359098 | -0.3407860 | 0.7332647 | 0.7258258 | 1.2376321 |
| europe-other | 1.0879700 | 0.1026601 | 0.8212882 | 0.4114821 | 0.8875542 | 1.3277526 |
| mena-other | 0.7727728 | 0.0957292 | -2.6927025 | 0.0070875 | 0.6396294 | 0.9311512 |
| multiple continents-other | 1.0004930 | 0.1815430 | 0.0027147 | 0.9978340 | 0.6886185 | 1.4061150 |
| north america-other | 1.2397122 | 0.0985847 | 2.1796415 | 0.0292840 | 1.0200138 | 1.5016587 |
| south america-other | 1.5318366 | 0.1748517 | 2.4390232 | 0.0147270 | 1.0700264 | 2.1278128 |

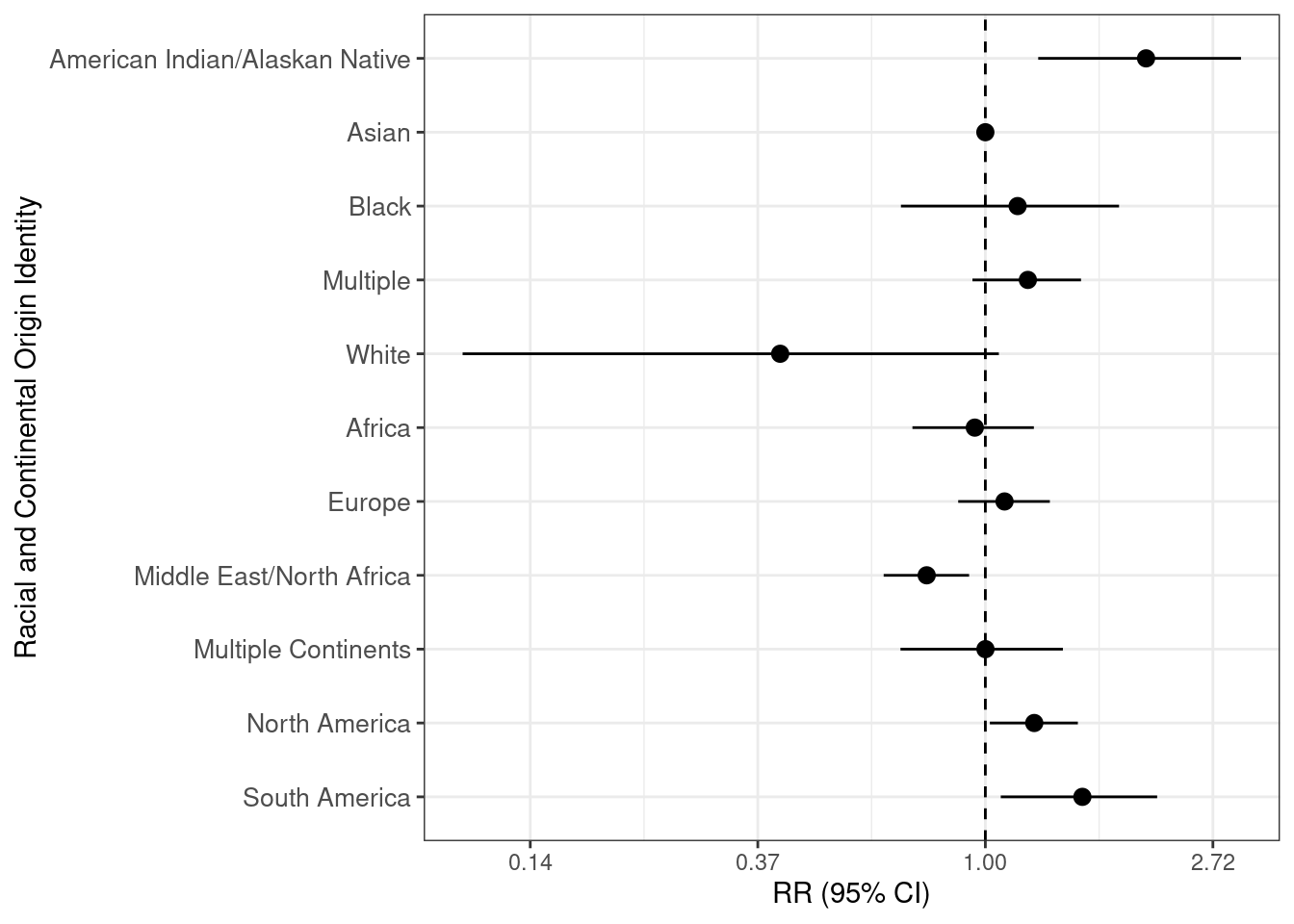
Hide

non\_hispanic\_poisson\_fp <- non\_hispanic\_poisson\_mod\_sum %>%

mod\_forest\_plot()

non\_hispanic\_poisson\_fp

## Warning: Removed 1 rows containing missing values (`geom\_segment()`).



Hide

ggsave("tables\_figures/non\_hispanic\_poisson\_forest\_plot.jpeg", non\_hispanic\_poisson\_fp, width = 8, height = 6, dpi = 300, units = "in")

## Warning: Removed 1 rows containing missing values (`geom\_segment()`).

## Chi-Square Test

To test if there is dependence in the distribution between preterm birth and the cleaned write-in variable, we perform a chi-square test. Since we have small bin sizes for some of the race categories, we will use a Monte Carlo simulated method. The way the Monte Carlo simulation works is to assume that the margins (summing across the race groups and the preterm birth categories) are fixed, then simulate contingency tables that would yield the same margins as our observed data. Then the chi-square statistic is calculated for each of these contingency tables, and we obtain a p-value by comparing the test statistic for our observed data and seeing what proportion of the simulated test statistics are as extreme or more extreme than the observed. If the p-value is below our prescribed level alpha of significance, then we reject the null hypothesis that there are no differences between classes in the population (i.e. independence between the categories of the two groups).

### Hispanic

Hide

pa\_data\_chisq <- pa\_data |>

filter(racem\_me\_mapped == "other") |>

filter(!race\_write\_in\_cleaned %**in**% c("native hawaiian/pacific islander", "oceania-other", NA))

with(

pa\_data\_chisq %>% filter(hispynM == "Hispanic"),

chisq.test(ptb, race\_write\_in\_cleaned, simulate.p.value = TRUE, B = 20000)

)

##

## Pearson's Chi-squared test with simulated p-value (based on 20000

## replicates)

##

## data: ptb and race\_write\_in\_cleaned

## X-squared = 27.966, df = NA, p-value = 0.0055

### Non-Hispanic

Hide

with(

pa\_data\_chisq %>% filter(hispynM == "Non-Hispanic"),

chisq.test(ptb, race\_write\_in\_cleaned, simulate.p.value = TRUE, B = 20000)

)

##

## Pearson's Chi-squared test with simulated p-value (based on 20000

## replicates)

##

## data: ptb and race\_write\_in\_cleaned

## X-squared = 39.422, df = NA, p-value = 2e-04

# Sankey Diagram

This Sankey diagram shows (moving left to right) how those who chose “Other” were sorted into the race categories on the birth certificates, then into race cateogires and continents of origin.

Hide

make\_sankey <- **function**(hispanic, include\_indeterminate) {

*# Make initial data clean statements*

sankey\_data <- pa\_data %>%

{

**if**(hispanic == "all") {.}

**else** **if**(hispanic == "hispanic") {filter(., hispynM == "Hispanic")}

**else** **if**(hispanic == "non-hispanic") {filter(., hispynM == "Non-Hispanic")}

} %>%

{

**if**(include\_indeterminate) {.}

**else** {filter(., race\_write\_in\_cleaned != "indeterminate")}

}

*# Make a connection data frame*

links1 <- sankey\_data |>

count(racem\_me\_mapped, racem\_me\_write\_in\_groups) |>

filter(racem\_me\_mapped == "other") |>

mutate(

racem\_me\_mapped = paste0(racem\_me\_mapped, "\_orig"),

racem\_me\_write\_in\_groups = paste0(racem\_me\_write\_in\_groups, "\_map1")

) |>

rename(**source** = racem\_me\_mapped, target = racem\_me\_write\_in\_groups)

links2 <- sankey\_data |>

count(racem\_me\_mapped, racem\_me\_write\_in\_groups, race\_write\_in\_cleaned) |>

filter(racem\_me\_mapped == "other") |>

select(racem\_me\_write\_in\_groups, race\_write\_in\_cleaned, n) |>

mutate(

racem\_me\_write\_in\_groups = paste0(racem\_me\_write\_in\_groups, "\_map1"),

race\_write\_in\_cleaned = if\_else(is.na(race\_write\_in\_cleaned), "no write-in\_map2", paste0(race\_write\_in\_cleaned, "\_map2"))

) |>

rename(**source** = racem\_me\_write\_in\_groups, target = race\_write\_in\_cleaned)

links <- rbind(links1, links2)

*# From these flows we need to create a node data frame: it lists every entities involved in the flow*

nodes <- data.frame(

name = c(as.character(links$**source**), as.character(links$target)) %>% unique()

) |>

arrange(

match(name,

c(

"other\_orig",

"Race\_map1",

"Continental Origin\_map1",

"other (no write-in)\_map1",

"american indian/native alaskan\_map2",

"asian\_map2",

"black\_map2",

"multiple\_map2",

"native hawaiian/pacific islander\_map2",

"white\_map2",

"africa-other\_map2",

"asia-other\_map2",

"europe-other\_map2",

"indigenous\_map2",

"mena-other\_map2",

"multiple continents-other\_map2",

"north america-other\_map2",

"oceania-other\_map2",

"south america-other\_map2",

"indeterminate\_map2",

"no write-in\_map2"

)

)

)

*# With networkD3, connection must be provided using id, not using real name like in the links dataframe.. So we need to reformat it.*

links$IDsource <- match(links$**source**, nodes$name)-1

links$IDtarget <- match(links$target, nodes$name)-1

*# prepare color scale: I give one specific color for each node.*

*# my\_color <- 'd3.scaleOrdinal() .domain(["group\_A", "group\_B","group\_C", "group\_D", "group\_E", "group\_F", "group\_G", "group\_H"]) .range(["blue", "blue" , "blue", "red", "red", "yellow", "purple", "purple"])'*

*# Remove suffix from source and target for easier to read labels*

strip\_suffix <- **function**(dat, cols) {

dat |>

mutate\_at(

cols,

**function**(x) gsub("(\_orig|\_map1|\_map2)$", "", x) |> substitute\_labels()

)

}

links <- strip\_suffix(links, c("source", "target"))

nodes <- strip\_suffix(nodes, "name")

*# Make the Network. I call my colour scale with the colourScale argument*

p <- sankeyNetwork(Links = links, Nodes = nodes, Source = "IDsource", Target = "IDtarget",

Value = "n", NodeID = "name", fontSize = 30, nodePadding = 25

*# colourScale=my\_color*

)

**return**(p)

}

## All patients with indeterminate included

Hide

make\_sankey(hispanic = "all", include\_indeterminate = TRUE)

OtherRaceContinental OriginOther (No Write-In)American Indian/Alaskan NativeAsianBlackMultipleNative Hawaiian/Pacific IslanderWhiteAfricaEuropeMiddle East/North AfricaMultiple ContinentsNorth AmericaOceaniaSouth AmericaIndeterminateNo Write-In

## Hispanic patients with indeterminate included

Hide

sankey\_hisp\_indet <- make\_sankey(hispanic = "hispanic", include\_indeterminate = TRUE)

sankey\_hisp\_indet

OtherRaceContinental OriginOther (No Write-In)American Indian/Alaskan NativeAsianBlackMultipleNative Hawaiian/Pacific IslanderWhiteAfricaEuropeMiddle East/North AfricaMultiple ContinentsNorth AmericaSouth AmericaIndeterminateNo Write-In

Hide

*#saveNetwork(sankey\_hisp\_indet, "output/sankey\_diagrams/sankey\_hisp\_indet.html")*

*# webshot::webshot("output/sankey\_diagrams/sankey\_hisp\_indet.html", "output/sankey\_diagrams/sankey\_hisp\_indet.png")*

## Non-Hispanic patients with indeterminate included

Hide

sankey\_nonhisp\_indet <- make\_sankey(hispanic = "non-hispanic", include\_indeterminate = TRUE)

sankey\_nonhisp\_indet

OtherRaceContinental OriginOther (No Write-In)American Indian/Alaskan NativeAsianBlackMultipleNative Hawaiian/Pacific IslanderWhiteAfricaEuropeMiddle East/North AfricaMultiple ContinentsNorth AmericaOceaniaSouth AmericaIndeterminateNo Write-In

Hide

*#saveNetwork(sankey\_nonhisp\_indet, "output/sankey\_diagrams/sankey\_nonhisp\_indet.html")*

*# webshot::webshot("output/sankey\_diagrams/sankey\_nonhisp\_indet.html", "output/sankey\_diagrams/sankey\_nonhisp\_indet.png")*

## Hispanic patients with indeterminate excluded

Hide

sankey\_hisp\_no\_indet <- make\_sankey(hispanic = "hispanic", include\_indeterminate = FALSE)

sankey\_hisp\_no\_indet

OtherRaceContinental OriginAmerican Indian/Alaskan NativeAsianBlackMultipleNative Hawaiian/Pacific IslanderWhiteAfricaEuropeMiddle East/North AfricaMultiple ContinentsNorth AmericaSouth America

Hide

saveNetwork(sankey\_hisp\_no\_indet, "output/sankey\_diagrams/sankey\_hisp\_no\_indet.html")

## Non-Hispanic patients with indeterminate excluded

Hide

sankey\_nonhisp\_no\_indet <- make\_sankey(hispanic = "non-hispanic", include\_indeterminate = FALSE)

sankey\_nonhisp\_no\_indet

OtherRaceContinental OriginAmerican Indian/Alaskan NativeAsianBlackMultipleNative Hawaiian/Pacific IslanderWhiteAfricaEuropeMiddle East/North AfricaMultiple ContinentsNorth AmericaOceaniaSouth America

Hide

saveNetwork(sankey\_nonhisp\_no\_indet, "output/sankey\_diagrams/sankey\_nonhisp\_no\_indet.html")

## Make Sankey diagram panel diagram

Hide

sankey\_hisp\_indet\_path <- file.path("output" , "sankey\_diagrams", "sankey\_hisp\_indet.png")

sankey\_nonhisp\_indet\_path <- file.path("output", "sankey\_diagrams", "sankey\_nonhisp\_indet.png")

p1 <- ggdraw() + draw\_image(sankey\_hisp\_indet\_path)

p2 <- ggdraw() + draw\_image(sankey\_nonhisp\_indet\_path)

plot\_grid(p1, p2, labels = c('A', 'B'))