Problem set 1: Question 1

Latera Tesfaye Olana

18 January, 2023

Improving plot from data triangulation project

The present figure was extracted from the work of Latera et al., entitled Estimating Immunization Coverage at the District Level: A Case Study of Measles and Diphtheria-Pertussis-Tetanus-Hib-HepB Vaccines in Ethiopia. Figure 1 illustrates vaccine coverage estimations for the first dose of Measles and the third doses of Pentavalent (diphtheria-pertussis-tetanus-Hib-HepB) across Ethiopian regions utilizing four distinct data sources. The first dataset, referred to as the Ethiopian Demographic Health Survey (EDHS), is considered to be the most accurate and serves as the gold standard. The second, third, and fourth data sources are Worldpop (WP), the District Health Information System (DHIS), hereafter referred to as the Health Information System, and the Central Statistics Agency (CSA), respectively. These three datasets comprise the population of under-five children (our targeted population) at the district level, with district being one administrative level below region (smaller). To estimate the vaccine coverage from these three data sources, hospital records containing the number of vaccines administered at the district level were utilized. The estimated vaccine coverage from these three data sources is generally referred to as administrative coverage. Due to reasons beyond the scope of this assignment, these three estimates are highly biased and tend to overestimate coverage at the district level. It was not possible to identify specific districts in which coverage was overestimated, underestimated, or accurate. As coverage at the regional level (similar to that displayed in Figure 1) is a weighted average of administrative coverage, the bias remains unchanged at regional levels. Since our model assumes equal expected bias across districts (for the three datasets), we cannot infer the quality of the model or data based on whether coverage estimation exceeds 100% or falls below 100%.

For the paper, we wanted to design a graph that could convey the following messages:

- Show the coverages of both vaccines from all the data-sets for all the regions
- Show how the estimation of WP, DHIS, and CSA varies from the gold standard (DHS)
- The convergence of WP, DHIS, and CSA might imply the gold standard is incorrect.

So, what is the graph doing right?

- The choice of the graph (scatter) and the design is clear
- It is easy to understand
- The design of the grid-lines and the vertical dashed lines linking the graph at the bottom with the one at the top
- Visually appealing. For instance, in order to enhance readability, the sequence of the four estimates for a specific region was arranged in a vertically offset manner.

Any room for improvement?

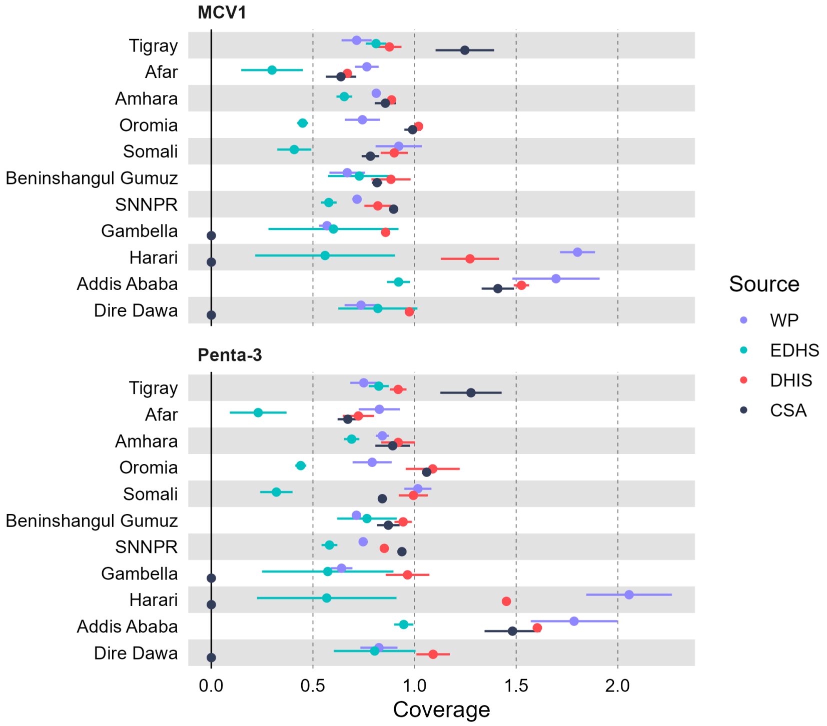

Figure 1: Vaccine coverage of the first dose of Measles and the third doses of Pentavalent across Ethiopian regions. (This plot where taken from Latera et al. paper - unpublished)

- The uncertainty plots presented for DHIS, CSA, and WP may not necessarily provide a reliable representation of the data. In the realm of statistical estimation, a narrower confidence interval is commonly perceived as a more precise measure. However, in this particular scenario, certain regions may have a small number of districts and the population numbers for those districts may have been overestimated or underestimated all together. For example, data from the WP tends to overestimate population numbers in urbanized regions. This implies that most districts in those regions will have large overestimated population numbers. However, due to the proximity of these high values, the standard error will be smaller, resulting in a narrower uncertainty range. In summary, the districts from which the regional weighted mean is calculated may have a biased population numbers (all of them badly high or low), but with a smaller standard error.

- The ordering of the legends seems random

- The level of uncertainty within the EHDS data-set is of scientific significance due to its nature as a survey data-set. The standard error within the sample directly impacts the level of uncertainty. However, in relation to the overall research objective, the level of uncertainty within the data-set does not provide any additional informative insights.

- The label coverage on the horizontal access doesn't have unit of representation (in this case %)

- The legend 'Sources' is not clear

- The legend texts, the vaccine types, and some regions are abbreviated or in short form. Thus difficult to translate

- In statistical analysis, it is common practice to stratify data across various characteristics within a given group. For example, the differentiation of individuals based on their assigned birth gender may hold scientific significance. However, in this particular instance, stratifying the data based on vaccine type is unnecessary. This is mainly due to we are not interested in analyzing the differential or comparative characteristics of these two vaccines. Such distinctions or similarities hold no significance and are not useful for inferences to be made about one vaccine based on the characteristics of the other. Plotting both vaccines side by side will not provide any scientifically useful information and the two graphs must be separated.

- The CSA (Central Statistical Agency) estimate for Dire Dawa, Harari, and Gambella regions shows zero. This is a result of our model excluding CSA's estimations for these three regions. Therefore, these regions should be designated as "missing" rather than "0", as the latter implies a specific numerical value.

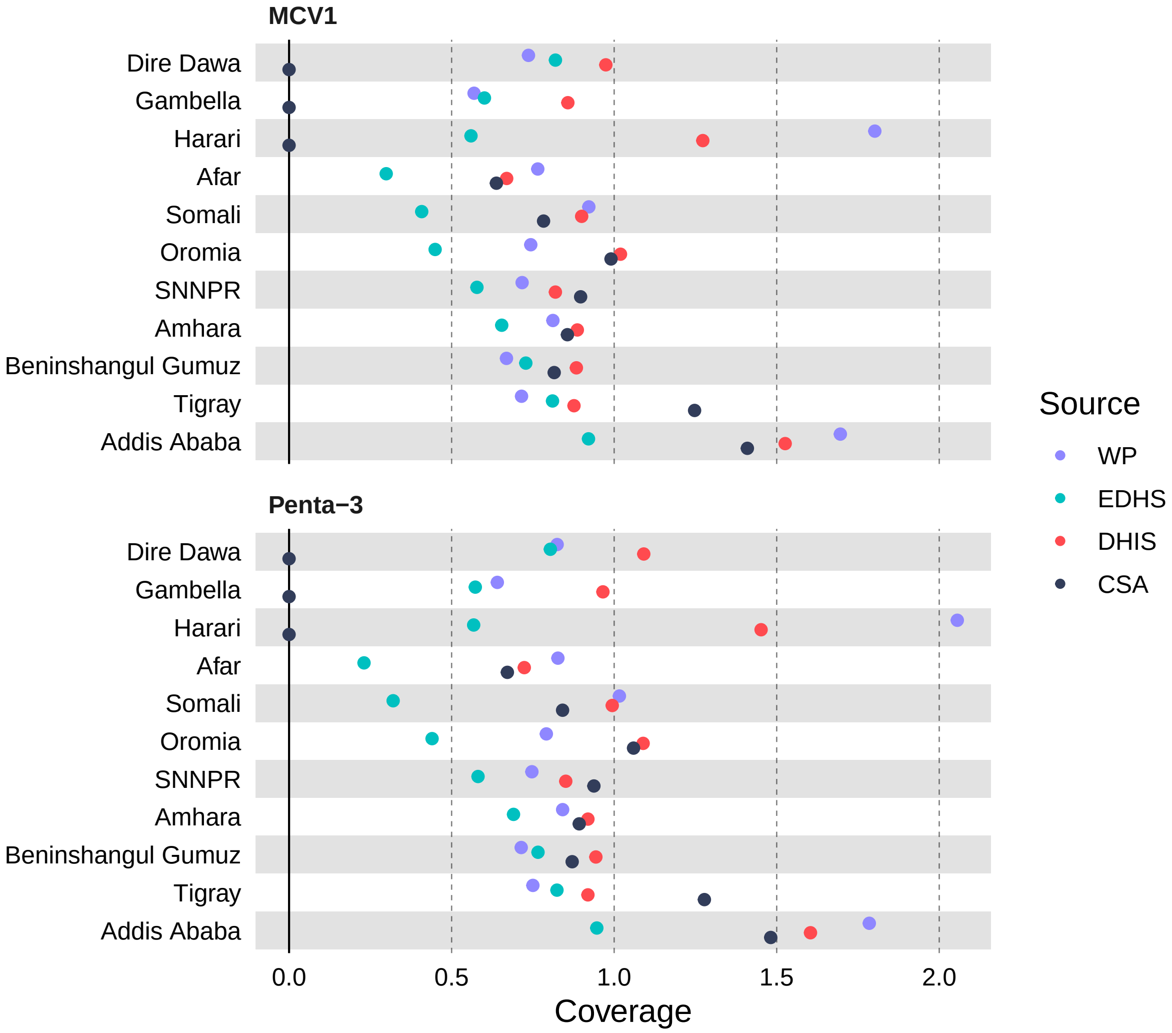

Figure 2: Standard error removed - Vaccine coverage of the first dose of Measles and the third doses of Pentavalent across Ethiopian regions. (This plot where taken from Latera et al. paper - unpublished)

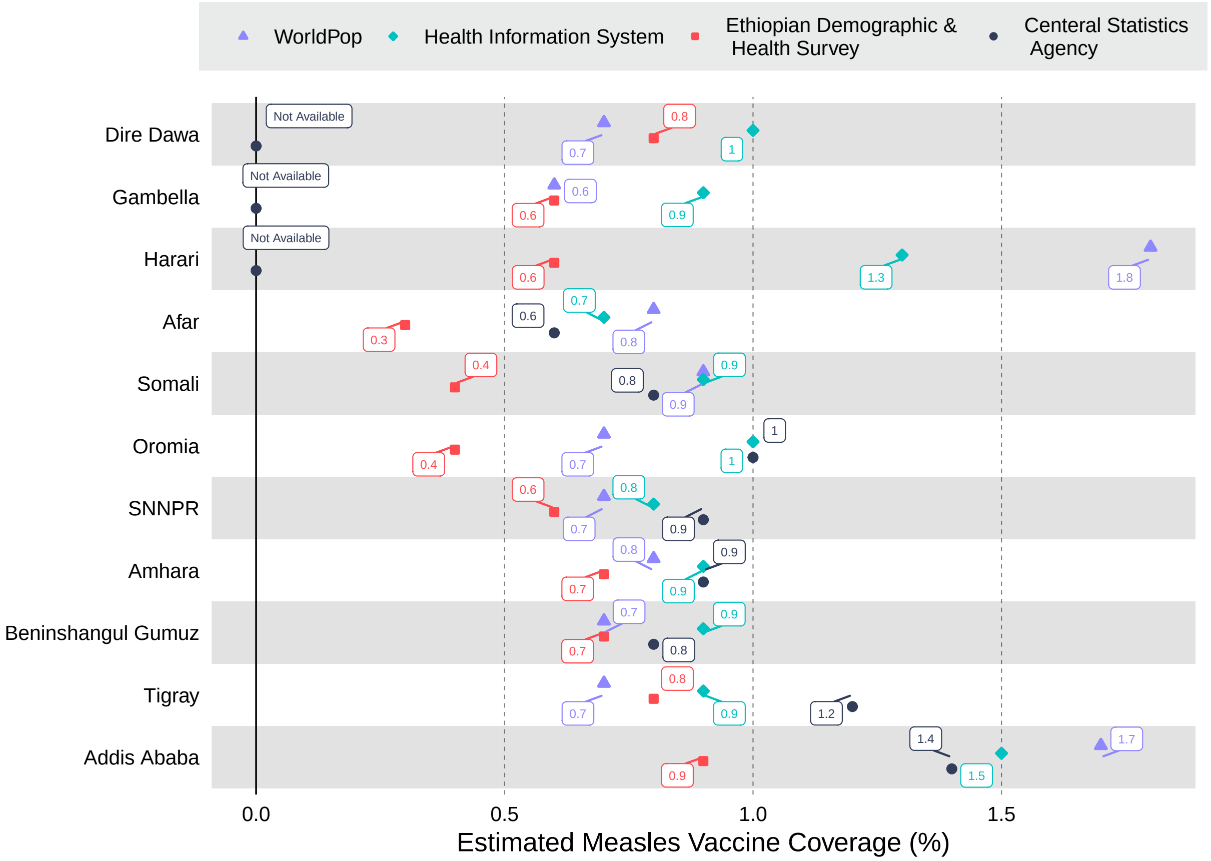

Figure 3: Improved - Vaccine coverage of the first dose of Measles across Ethiopian regions for WorldPop, Ethiopian Demographic & Health Survey, Centeral Statistics Agency and, Health Information System data-sources. Note: SNNPR: South Nations Nationality and People

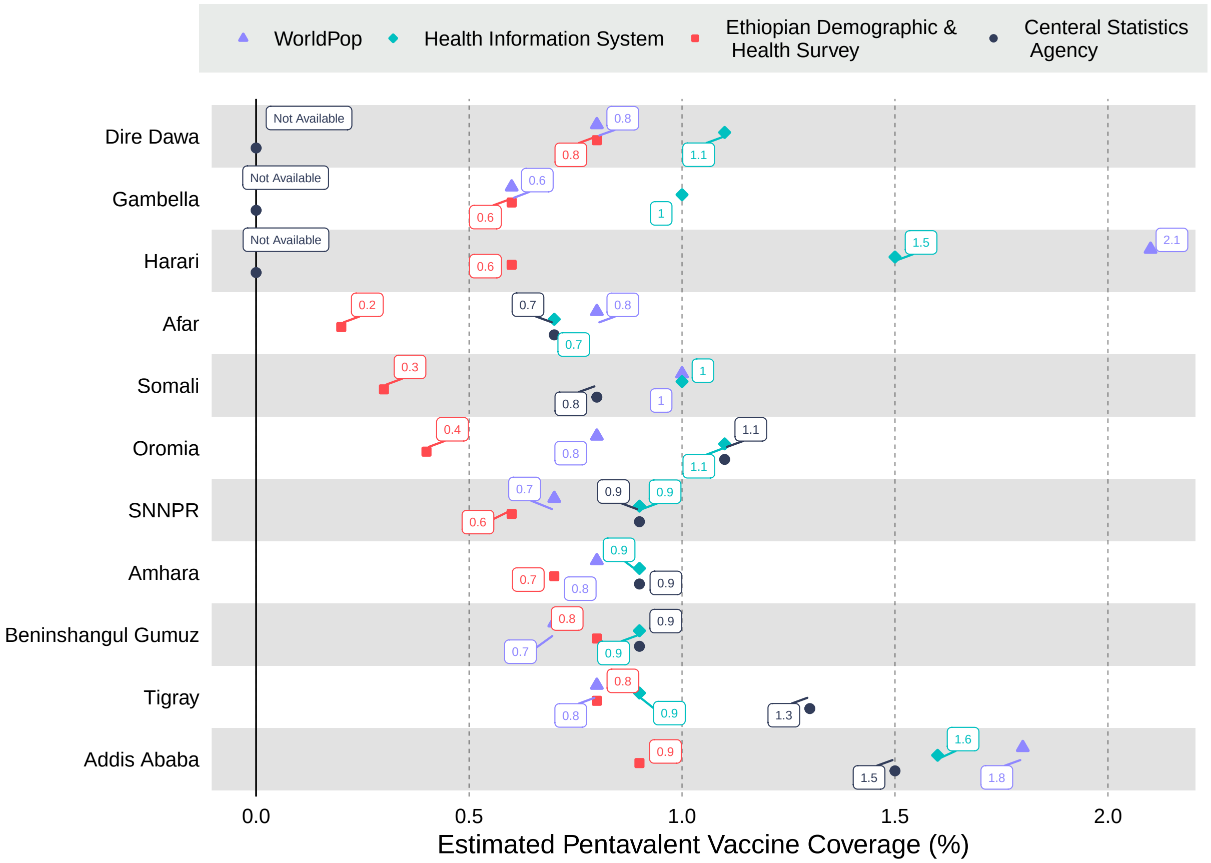

Figure 4: Improved - Vaccine coverage of the third doses of Pentavalent across Ethiopian regions for WorldPop, Ethiopian Demographic & Health Survey, Centeral Statistics Agency and, Health Information System data-sources. Note: SNNPR: South Nations Nationality and People

Code Appendix

```
### Setting up the packages
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
# check if packages are installed; if not, install them
packages <- c("dplyr", "readr", "ggExtra", "plotly",
              "ggplot2","ggstatsplot","ggside","rigr","nlme","lmtest",
              "sandwich","hrbrthemes","MASS","magrittr","ggforce","ggrepel")
not_installed <- setdiff(packages, rownames(installed.packages()))
if (length(not_installed)) install.packages(not_installed)
#install.packages("devtools")
#devtools::install_github("NightingaleHealth/ggforestplot")
# load packages
library("MASS")
library(dplyr)
library(sandwich)
library(readr)
library(lmtest)
library(ggforestplot)
library(nlme)
library(magrittr)
library(ggforce)
library(ggrepel)
library(ggstatsplot)
library(ggside)
library(rigr)
library(hrbrthemes)
library(ggExtra)
library(plotly)
library(ggplot2)
# library(tidyverse) # don't load tidyverse package due to conflict with dplyr
### -----------------------------------------------------------
#Loading working directory of the raw data

#Please load your data/directory by changing it with your work directory
#Throughout this code module you will see a tone of places, where
#data is read and written, so please make sure to change them to your
#working directory folder format

working_directory_data <- setwd("C:/Users/latera/Desktop/viz/")

raw <- readxl::read_xlsx("data/final_data.xlsx")
knitr::include_graphics("final_plot_viz.png")

#Remove SE by changing value to zero, we can't delete it
#(as ggforest won't work other wise)
raw$se <- 0
ggforestplot::forestplot(df = raw,
                         name = name,
                         estimate = Coverage,
                         colour = Source,
                         logodds = FALSE) +
  ggforce::facet_col(
   facets = ~ groups,
   scales = "free_y",
   space = "free"
 )

#Add correct labeling
#Change the ordering of grouping (alphabetically)
raw$se <- 0

#Change name of legends.
raw <- raw %>%
  mutate(vaccination_name = case_when(groups == "MCV1" ~
                                        "First dose of measles",
                                      groups == "Penta-3" ~
                                        "Third doses of pentavalent"),
         data_source_name = case_when(Source == "WP" ~ "WorldPop",
         Source == "EDHS"~"Ethiopian Demographic & \n Health Survey",
         Source =="CSA"~"Centeral Statistics \n Agency",
         Source =="DHIS"~"Health Information System"))

#If not rounded the precision points will cluster the plot
raw <- raw %>%
  mutate(Coverage = round(Coverage, 1))

#filter the data only for MCV1
raw_mcv1 <- filter(raw, groups == "MCV1")

#Change zero value to 'Un-avaliable'
raw_mcv1<-raw_mcv1 %>% mutate(Coverage_new = ifelse(Coverage == 0,
                                      "Not Available", Coverage))

#Adding labels on each points.
ggforestplot::forestplot(df = raw_mcv1,
                         name = name,
                         estimate = Coverage,
                         colour = data_source_name,
                         shape = data_source_name,
                         xlab='Estimated Measles Vaccine Coverage (%)',
                         logodds = FALSE) +
  ggplot2::theme(legend.position = 'top')+
  theme(legend.background = element_rect(color=NA,
                                         fill = "#e8ebe9"))+
  theme(legend.title = element_blank())+
  geom_label_repel(aes(label=Coverage_new,
                       color=data_source_name),
                   size=2,max.overlaps =
                     getOption("ggrepel.max.overlaps",
                               default = 40),show.legend = F)



#filter the data only for MCV1
```

```r
raw_dpt3 <- filter(raw, groups == "Penta-3")

#Change zero value to 'Un-avaliable'
raw_dpt3<-raw_dpt3 %>% mutate(Coverage_new = ifelse(Coverage == 0,
                                      "Not Available", Coverage))
#Adding labels on each points.
ggforestplot::forestplot(df = raw_dpt3,
                         name = name,
                         estimate = Coverage,
                         colour = data_source_name,
                         shape = data_source_name,
                         xlab='Estimated Pentavalent Vaccine Coverage (%)',
                         logodds = FALSE) +
  ggplot2::theme(legend.position = "top")+
  theme(legend.background = element_rect(color=NA,
                                         fill = "#e8ebe9"))+
  theme(legend.title = element_blank())+
  geom_label_repel(aes(label=Coverage_new,
                       color=data_source_name),
                   size=2,max.overlaps =
                     getOption("ggrepel.max.overlaps",
                               default = 40),show.legend = F)
```