EPIB 704 - Semester-wide Assignments for Fall 2025

Instructor: Mabel Carabali

Assigned: 2025-09-11 (Last Assignment due: November 20th, 2025)

Contents

Premises. 1
Setup. 2
Assignment 1. [20 Points Total = 15 Points for questions + 5 for style] 4

Premises.

This document contains the Homework assignment for the EPIB 704 Fall 2025 course. The total score is **100 points**. Grading for each of the five sub-assignment is out of **20 points total**, with **15** points from the questions and 5 points for overall style, organization, and clarity.

Please note that 100 points correspond to **75%** of the overall EPIB 704 grade. Recall that each one of the five assignment has the same weight **(15%)** of the overall grade.

Please show your work (including software functions and other calculations) and format your responses appropriately (e.g. a reasonable number of digits and rounding only when necessary, 0.5 points per question will be deducted if inappropriate or incorrectly rounding).

Schedule

The schedule of dates for when these will be assigned, due (handed in) and returned (handed back) is as follows:

HW #	Assigned	Due	Returned
1	Sept 11	Sept 18	Sept 25
2	Sept 25	Oct 02	Oct 09
3	Oct 09	Oct 23	Oct 30
4	Oct 30	Nov 06	Nov 13
5	Nov 06 *	Nov 20	Nov 27

* Last assignment is given 15 working days in advance. See Policy on Assessment of Student Learning (PASL) subsection 6.7. in Point 6. Communication of course assessment tasks.

Howework Grading

Some collaboration in homework assignments may be beneficial but please use good judgment in preventing your collaboration from becoming detrimental to your learning of the material. Submitted assignments should be your individual effort, even if you consult with other students about your strategy for obtaining these solutions (see plagiarism note below).

To reinforce the concept of scientific **reproducibility** all assignments should be submitted as `R-Markdown` files. Data files for the assignments are available on GitHub and may be installed directly into `R` . **Details in the GitHub tab.**

Assignments should be submitted before 23:59 (11:59 PM) on the due date.

- Late assignments carry a very severe penalty of **20% off per day late**. This is primarily to protect your time, help you with time management and the time of the TAs.
- Pleas of mercy for extenuating circumstances will be accepted only with written documentation. Taking extra time to do a better job is probably not a worthwhile strategy, because the late penalty is so costly.

EPIB 704 GitHub Repository

Data for assignments can be found directly in the `/EPIB-704/tree/main/data` folder. And we have three options for you to import the assignments datasets into R:

1. **Option 1:** Install our 'epib.704.data' package:

a) First, install the remotes package: install.packages("remotes")

b) Then, install the 'epib.704.data' package: remotes::install_github("mcarabali1/epib.704.data")

c) Load the dataset, which can be identified by typing `epib.704.data::`

2. **Option 2:** Install them directly in R using their URL:

a) Assign the dataframe (df)) URL to an R object: Using the following code sintax:

```
urlfile <- "https://raw.githubusercontent.com/mcarabali1/EPIB-704/main/data/covidkenya.csv"
```

b) Import the dataframe `df` into R: Using the following code sintax

```
mydata <- read.csv(url(urlfile))
```

3. **Option 3:** Download the files directly from the repository onto your computer and import them into R.

Use of Generative AI (Artificial Intelligence) tools:

AI generative tools include but are not limited to ChatGPT $^{(R)}$. There is an existing dialogue on policies about the use of these AI generative tools. Extreme care is recommended in the use of generative AI as a learning tool in the context of EPIB 704 in general and especially for the homework assignments. Each assignment has a pedagogical objective and is intended to solidify your knowledge and enhance the learning experience. My goal is that, at the end of the term you can perform simple calculations or estimations without the need to rely on AI tools. It is, and should be more about knowing where the estimations come from and how they are made.

Although the use of ChatGPT and any other generative AI tool cannot be discouraged, and understanding that AI tools in general are part of the future of the epidemiological profession; I expect that if you do use any if these tools, please abide by the principle of transparency and inform, disclose or report the type of tool, software, and prompt (or combination of prompts) used to obtain you answer. As you may realize throughout the term, even if these tools help you to obtain a numerical answer, these tools may not help you in getting the expected answer. You are expected to provide interpretations and comments reflecting your critical thinking process, which is expected to come from you and not from these tools.

As per PASL, *McGill has approved a secure version of Copilot for use within the University. We encourage you to review these guidelines on using gen AI at McGill and this resource on using gen AI in teaching and learning.*

Setup.

Before we start, load some useful packages including the data for this course `epib.704.data`, for data manipulation and plotting, the `tidyverse` package which includes `dplyr` and `ggplot2` are useful. For simple

epidemiology regressions we will be using `glm` and `jtools` for data analysis, and may need `episensr` for bias correction.

NOTE:

1) Click on "show" to see the packages included here. Feel free to add other packages as you learn/consider it relevant;

2) Make sure both `R` and `RStudio` are up to date in your computer. Check Here for R and here for R-Studio

3) Please use the `set.seed` for reproducibility. Recommend using `set.seed(7042025)`.

```
set.seed(7042025)
packages <- c("tidyverse", "ggdag", "dagitty",  "here", "knitr", "patchwork", "brms", "stdReg",
              "ggridges", "gmodels", "xaringanthemer", "latex2exp", "epiR", "margins","dplyr",
              "rstanarm","table1", "pvaluefunction", "survival", "cobalt", "survminer","logbin",
              "broom", "ggdist", "jtools", "knitr", "GGally", "table1", "Epi","glm2", "tidyverse",
              "epitools", "ggplot2", "brms", "rstanarm","cmdstanr", "epib.704.data",  "gtsummary",
              "rstan", "ggstats", "bayestestR", "see", "margins", "boot", "interactionR","latex2exp",
              "interactions",   "MatchIt","WeightIt", "cobalt", "episensr", "kableExtra", "epibasix",
              "epiDisplay", "lmtest", "sandwich", "foreign", "ipw", "survey", "cmdstanr",
              "ggpubr", "gt", "rmdformats", "prettydoc")


invisible(lapply(packages, function(xxx)
  suppressMessages(require(xxx, character.only = TRUE,
                           quietly=TRUE, warn.conflicts = FALSE))))
```

Resources

For this assignment we are going to use data available from recent scientific publications. The three scientific manuscripts and respective associated data are presented below. If you feel compelled to contact the authors to get additional information, I won't be opposed. Just please consider this is an academic exercise and objective of the homework may differ from the scientific question posed by the authors and hence their methodological approach.

Dataset #ID	Dataset name	Related article title
1	"Amazonas_HQoL"	Andrade MV, et al. (2024) Health-Related Quality of Life due to malaria in the Brazilian Amazon using EQ-5D-3L. PLOS Neglected Tropical Diseases 18(12): e0012739.
2	"Covid_Bangladesh"	Islam MN,et al. (2025) High coverage and equitable distribution of COVID-19 vaccine uptake in two vulnerable areas in Bangladesh. PLOS Global Public Health 5(1): e0004178.
3	"VL_Nigeria"	Obasa GB, et al. (2024) Factors associated with viral load re-suppression after enhanced adherence counseling among people living with HIV with an initial high viral load result in selected Nigerian states. PLOS Global Public Health 4(11): e0002876.

Data Assignment

Below you have the list of your names with the respective data assignment. All information is included here, attached, as hyperlink or in the GitHub repository in the folder: EPIB704_HW_data_2025. I randomly

Datasets assignment

EPIB 704 Homework assignments -FALL 2025

nb	McGill ID	Dataset ID
1	260859889	2
2	261146963	2
3	260919502	3
4	261264774	1
5	261019649	1
6	261263833	2
7	260896136	2
8	261072609	1
9	261271369	3
10	261271256	3
11	261270629	1
12	261151122	3
13	261213800	1
14	261265010	3
15	260742427	2

Please find ancillary documentation on GitHub

assigned the documents to each of you. The list is presented here and was obtained using the following code `classlist2025$db<- ceiling(runif(length(classlist2025$nb), min=0, max = 3))`.

Assignment 1. [20 Points Total = 15 Points for questions + 5 for style]

1. Critical appraisal [This section is not graded]

Read the manuscript and assess the supplementary information to answer the following:

a. What's the presented research question and overall objective of the study.

b. A *priori* do you think the study design and methodological approaches are adequate to answer the research question? (provide a brief explanation more than three sentences).

c. Consequently, do you think the result and conclusions are consistent with the objectives and methods? Hence, did the author **answered** their research question?.

2. Data management [3 Points Total]

Following the instructions on the premises and setup, upload the data assigned to:

a. Revise the structure and dimension of your dataset. Check whether the sample size in your dataset correspond to the one reported by the authors. [Hint: can use `str` or `dim` functions] [1 point]

b. Generate a "Table 1" using the information provided in the manuscript. This is, if the manuscript has a Table 1, reproduce it with the data at hand. If your manuscript does not have it, then generate a summary table, providing measures of frequency and dispersion according to the type of variable (e.g., IQR, Inter Quartile Ranges). The table should include minimum and maximum values and for continuous variables and proportions for categorical variables, as well "NAs", missing, and or "unknonw values". [Hint: Use any R-package, e.g. `gtsummary` with the `tbl_summary` function to reproduce/generate a Table 1 with the given covariates] [1 point]

Data-Specific Note: for the `VL_Nigeria` data, omit the time-to-event variables (e.g., Time to complete EAC sessions, Time from EAC session completion to repeat viral load (VL)).

c. Comment on any consistency or lack of between the reproduced tables and the ones published or reported in the manuscript, if any. [1 point]

3. Research questions and DAGs [4 Points Total]

Using the information form the manuscript, identify the "Exposure" or "Treatment", "Outcome", and the set of covariates included in the analysis/dataset. **Data-Specific Note:** for the `Covid_Bangladesh` data, use the variable `chroill_1` as the main exposure.

a. Provide a list of the PICO / PICOT criteria in you dataset and case. [0.5 point]

b. Draw a DAG for the relationship between exposure of interest and outcome of interest, including the main covariates listed in the manuscript (namely those included in their Tables 1 or 2). [1 point]

c. Given the DAG that you drew in Q.3.b, identify and list at least three structures that affect the relationship between "exposure" and "outcome". For example, identify "open back door paths,"blocked paths", colliders or confounders. Then, indicate what do you think that will be the "minimal" adjustment set to identify a causal association between "exposure" and "outcome"? Please try to answer this question without the aid of any software. [1 point]

d. Using the functions of `ggdag`, `dagitty` or any other `R-package`, and assuming this DAG is the correct DAG, and that we want to obtain a causal effect; identify what are the paths present in this relationship, what are the open paths, and what is the software-recommended adjustment set to identify an association between "exposure" and "outcome". Do you confirm what you wrote in Q.3.c? [Hint: What are the variables needed to be controlled for to determine the causal effect] [1.5 points]

4. Measures of Occurence and Association [8 Points]

Using the information in the manuscripts and the dataset itself, omitting clustering, matching, time-to-event, or any other structures in the dataset, consider the entire data as "baseline" information to answer the following:

4.1. Estimate and present the measures of occurrence for "outcome" and "exposure" and comment on the results. Report whether your estimates coincide with those presented in the manuscript. [Hint: It can be a prevalence or an incidence, you can use your Table 1 from Q.2.b or calculate it otherwise] [1 point]

4.2. Some of the manuscripts are presenting measures of association between the exposure of interest and outcome of interest.

a. Generate a variable representing the outcome of interest as dichotomous as follows: [1 point]

Dataset	Original variable	New Outcome	Notes: criteria to create the *new* outcome
"Amazonas_HQoL"	`P77` (VAS, visual analog scale for QoL)	`binQoL`	= 0 if VAS<=50; or = 1 if VAS >50
"Covid_Bangladesh"	`vdoses` (Taken at least two doses of COVID-19 vaccine)	`binVax`	= 0 if `vdoses` is 1 dose or less; or = 1 if `vdoses` is 2 or more doses.
"VL_Nigeria"	`VL Result` (Viral Load)	`binSupress`	= 1 if `VL Result` >=1000 copies/mL; or = 1 if `VL Result` <1000 copies/mL

b. Estimate the univariate, "crude" or unadjusted Risk Difference (RD), Risk Ratio (RR), and Odds Ratio (OR) for the association between the exposure of interest and outcome of interest. Interpret each of these measures and comment on whether these are consistent or different from those reported in the manuscript (if any) and why could these be same or different from the ones you estimate. [3 points]

4.3. Just for fun, conduct a null hypothesis testing for the association of exposure and outcome, via the estimation of an OR and a RD (Can take the ones estimated in 4.2.b).

a. Provide/Calculate/Obtain a *p-value* estimate for each of the measures of association between exposure and outcome and interpret them, to the best of your knowledge. [0.5 point]

b. Estimate/Provide the confidence interval for for each of the measures of association between exposure and outcome and interpret them, to the best of your knowledge. [0.5 point]

c. Plot the p-value function for each of the measures of association between exposure and outcome and interpret them, to the best of your knowledge. [Hint: can use software assisted plots with the codes provided on the slides, or the `pvaluefunction` R-package] Does your interpretation matched the one in Q.4.3a? [2 points]

Recall: Please note that 100 points correspond to **75%** of the overall EPIB 704 grade. Recall that each one of the five assignment has the same weight **(15%)** of the overall grade.
Please show your work (including software functions and other calculations) and format your responses appropriately (e.g. a reasonable number of digits and rounding only when necessary, 0.5 points per question will be deducted if inappropriate or incorrectly rounding).
