Assignment 1

Randomized Trials, Trial Emulation, and Data Collection (Weeks 1–2)

Ashley I Naimi

Due: Tuesday, September 15, 2026

1 Overview

This assignment covers material from Week 1 (Randomized Controlled Trials and Emulation) and Week 2 (Data Collection). It is worth **5 points**, allocated for completion, and is due **Tuesday, September 15**. You may resubmit as many times as you like over the course of the semester. You are encouraged to work in groups, but each of you must submit your own work.

The assignment consists of the free response questions below.

2 Free Response Questions

Answer each of the following. Where a question has parts, answer every part. Show your reasoning; a correct conclusion with no argument behind it will not receive useful feedback from your peer reviewer. Questions 1–5 cover Week 1; Question 6 covers Week 2.

2.1 Question 1: Response types and what exchangeability delivers

Following Greenland and Robins (1986), consider a closed cohort of N_1 treated and N_0 untreated individuals, all disease-free at baseline, followed over a fixed risk period. Let p_j and q_j ($j = 1, 2, 3, 4$) denote the proportion of the treated and untreated groups, respectively, belonging to response type j, where the types are: (1) doomed, (2) causative, (3) preventive, and (4) immune.

a) Show that $R_1 - R_0 = (p_1 + p_2) - (q_1 + q_3)$, and use this expression to explain why observing $R_1 - R_0 > 0$ is not sufficient to conclude that the treatment caused disease in anyone.

b) State the equality across groups that would license a causal reading of the risk difference. Explain what it means for that equality to hold "in expectation" under randomization, and why that is a claim about the design rather than about the dataset in front of you.

c) A colleague argues that because randomization delivers this equality, a trial's risk difference can be read as the proportion of participants harmed by treatment. Explain what is wrong with this.

2.2 Question 2: What randomization delivers beyond exchangeability

Consider two studies of the effect of adiposity on blood pressure. **Study A** is a Mendelian randomization analysis using variants in *FTO* and *MC4R* as instruments for BMI in a population-based cohort. **Study B** is a hypothetical randomized trial assigning participants to a 12-month weight-loss intervention or to usual care.

a) Grant Study A's central claim that the random segregation of alleles at meiosis renders exchangeability true. Name **three** distinct things Study B's investigators control by physically assigning treatment that Study A's investigators do not, and for each, name the specific threat to validity that the control addresses.

b) In an ideal randomized trial, the intention-to-treat contrast is characterized by exchangeability, positivity, and consistency by design. Explain what supplies each of the three. Then identify which of the three is most likely to be problematic in Study A, and defend your choice.

2.3 Question 3: Time zero, alignment, and immortal time

An investigator uses an insurance claims database to emulate a trial comparing two second-line agents for type 2 diabetes. Eligible patients are those with a type 2 diabetes diagnosis code between 2015 and 2018 who were receiving metformin monotherapy at diagnosis. The exposure is defined as the second-line agent the patient was receiving **12 months after the diagnosis date**. Follow-up begins at the **diagnosis date**. The outcome is all-cause mortality.

a) Identify the protocol component that has been violated, and state the rule that has been broken.

b) Name the resulting bias and describe its mechanism. Be specific about which person-time is misclassified, for whom, and in which direction the effect estimate is expected to move.

2.4 Question 4: Censoring, truncation, and competing events

a) Classify each of the following. For each, state which of right/left/interval censoring, right/left/interval truncation, or competing event applies, and say in one sentence why.

   i. A retrospective study of the time from receipt of HIV-contaminated blood products to AIDS diagnosis, in which individuals could be identified for enrollment only if they had already received an AIDS diagnosis.

   ii. A study of spontaneous abortion recruiting women at their first prenatal care visit, in a setting where early losses are frequently mistaken for a late menstrual period.

   iii. A five-year cohort study that records whether each participant developed hypertension by the study's end, but not the date of onset.

   iv. A participant in a 36-month trial who withdraws consent at month 14.

   v. A participant in a trial of a drug for prostate cancer mortality who dies of a myocardial infarction at month 20.

b) Explain the structural difference between right censoring and a competing event. Then explain why coding a competing event as a censoring event changes the estimand rather than simply biasing an estimate of a fixed target.

c) A colleague tells you their analysis has "no censoring problem," because the outcome was ascertained for everyone at the end of follow-up. Explain what type of censoring they nevertheless have, and what it costs them analytically.

2.5 Question 5: Intention-to-treat, per protocol, and as treated

Consider the single-time-point aspirin trial from the Week 1 notes. Let X denote randomization to aspirin ($X = 1$) or placebo ($X = 0$); let A denote the treatment actually taken: took aspirin ($A = 1$) or took placebo ($A = 0$); let C denote a confounder of the relationship between A and Y; and let Y denote the outcome. The structure is the one in the Week 1 notes: $X \to A \to Y$, with $C \to A$ and $C \to Y$.

a) Using the diagram, explain why $E(Y \mid X = 1) - E(Y \mid X = 0)$ is identified without any covariate adjustment, and state precisely, in words, what causal quantity it represents.

b) Two further contrasts are commonly reported:

$$E(Y \mid A = 1) - E(Y \mid A = 0) \qquad \text{and} \qquad E(Y \mid X = 1, A = 1) - E(Y \mid X = 0, A = 0)$$

Explain why neither is identified as a causal effect, being explicit about the path that is open in each case. State what would be required to estimate the per-protocol effect properly.

c) A published trial emulation reports a "modified intention-to-treat effect." Explain why this label is misleading, and state what the analysis is in fact estimating and under what assumptions.

2.6 Question 6: The hierarchy of information

Three teams each ran a five-year closed cohort study of the same outcome. All three enrolled everyone at a common time zero, and no one was lost to follow-up or experienced a competing event. The teams differ only in what they recorded:

- **Team A** recorded, for each participant, whether the outcome had occurred by the end of year 5 (yes/no), and nothing else.
- **Team B** recorded whether the outcome occurred *and* the exact time of each occurrence.
- **Team C** recorded only two totals for the whole cohort: the number of outcome events, and the total person-time accrued.

a) For each team, state which of the following are computable from their data, with a one-sentence justification for each: the 5-year risk; the risk function $F(t)$ over the full five years; the 5-year odds; the incidence rate over the five years; the median event time among those with events.

b) Team A's analyst claims their data are just as good as Team B's, "because the 5-year risk is all anyone needs." Give two scientific questions Team B can answer that Team A cannot, and connect your answer to why the Week 2 notes place the risk *function* at the top of the hierarchy of information.