Discussion Prompts: Designing a Study for a Novel Gene Variant

Scenario: Imagine a large-scale Genome-Wide Association Study (GWAS) has identified a novel gene variant, *VAR-X*. Preliminary data suggests *VAR-X* is significantly associated with accelerated cognitive decline in patients already diagnosed with Alzheimer's disease. As a team of epidemiologists and bioinformaticians, your group has been tasked with designing the definitive study to validate this association and understand its clinical impact.

Question 1: The Foundational Design Choice

Propose two distinct epidemiological study designs that could rigorously test the hypothesis that *VAR-X* is associated with accelerated cognitive decline in Alzheimer's patients. For this discussion:

- **Part A:** Choose one observational design (e.g., prospective cohort, retrospective cohort, or case-control) and one experimental design (e.g., a hypothetical randomized controlled trial for a therapeutic targeting the *VAR-X* pathway).
- **Part B:** For each design, define your specific study population, primary exposure, and primary outcome measure.

observational design: Prospective cohort

- study population: Diagnosed AZ patient with Var-X and without Var-X equally
- primary exposure: Presence of the VAR-X gene variant
- primary outcome: rate of cognitive decline

Experimental Design: Var-X therapy

- study population: Diagnosed AZ patient with Var-X
- primary exposure: therapy (treatment and placebo)
- primary outcome: change in rate of cognitive decline between treatment group and placebo group over a defined follow-up period

- **Part C:** Debate and justify which design would provide the strongest evidence for causality. What are the primary trade-offs between the two in terms of ethics, cost, feasibility, and time?

terms	Prospective Cohort	Var-X therapy
ethics	mostly concerned on collecting data, data accuracy, data privacy.	concerned similarly with prospective cohort with more complicated requirement for ethical experiment as it directly involve with patient.
cost	lower cost per person as it's not required control environment or intervention.	higher cost per person because higher cost due to drug development.
time	Longer duration: Alzheimer's patients are typically elderly, increasing risk of loss to follow-up (information bias).	If the process begins at the production, it may require approximately the same amount of time as the observation. shorter duration, if it starts from the experimental.
feasibility	it can be done successfully but may take a long period of time and not guarantee that the outcomes will be as expected.	The treatment may be ineffective (no improvement), making it impossible to find any association.

Question 2: Anticipating and Mitigating Bias

Focusing on the **observational study design** you proposed in Question 1, critically evaluate the potential threats to your study's validity.

- **Part A:** Identify and explain the two most likely and impactful sources of bias (e.g., selection bias, information bias/misclassification, confounding).

information bias - high overall **attrition** because of the advanced age of Alzheimer's patients; and the possibility of **information bias** from relying on interview data from participants.

confounding - interrupt both Var-X and Cognitive decline such as a**ge**, older patients may have both higher decline. **lifestyle factors**, such as smoking or diet, which affect neurodegeneration risk.

Part B: Propose specific, concrete strategies you would implement during the study design and data analysis phases to minimize the impact of these biases and any key confounders (e.g., age, disease severity at baseline, co-morbidities).

information bias: Increased frequency of follow-up and we might perform interviews along with other health records. According to interview, it should set cognitive decline standards to assess them on the same page; other health records are employed to validate the findings from the interview.

Confounding: Randomize and Matching: Match VAR-X positive and negative participants on key confounders such as **age**, **sex**, **baseline cognitive function** to ensure comparability between groups.

Question 3: The Bioinformatics & Big Data Perspective

Now, let's re-examine this research question through the lens of medical bioinformatics.

- Part A: How could you leverage existing large-scale resources, such as a national biobank or a federated network of Electronic Health Records (EHRs), to conduct a large-scale retrospective cohort study?

  Collect: aggregating Relevant Data --> Clean: ensuring Data Quality and Consistency --> Transform: structuring the Dataset for Analysis --> Analyze: performing Statistical

- Part B: Compare and contrast the internal and external validity of this "big data" approach versus a traditional, prospectively recruited cohort. What do you gain in terms of statistical power and generalizability, and what new challenges related to data quality and phenotype accuracy do you face?

Traditional cohorts ensure high internal validity with precise cause-and-effect control, while big data provides strong external validity through its large, diverse scale. In terms of **statistical power**, it helps detect small effects, and generalizability, it helps gaining ability to apply the results in a diverse population where small sample groups represent the population.

Challenges related to data quality: Data from big data could be inconsistent, different in format, inaccurate, incomplete data, which require more time to collect and process.

Challenges related to data phenotype accuracy: Difficulty in defining what is the phenotype that is supported, related to the progression of AZ disease.