Saranyaporn Kunawongkrit

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).

  ANS For the observational design, we chose a **prospective cohort study** to validate whether VAR-X is significantly associated with accelerated cognitive decline in Alzheimer's patients by observing and following those with and without the variant and comparing their respective rates of cognitive decline.

  We chose a **randomized controlled trial** for the experimental design. Alzheimer's patients will be randomly assigned to receive a treatment targeting VAR-X or a placebo, which will be added to their regular care. The study is to compare the rate of cognitive decline between the two groups.

- **Part B:** For each design, define your specific study population, primary exposure, and primary outcome measure.

  ANS

	Prospective cohort study	Randomized controlled trial
Study Population	Alzheimer's patients, equally divided by VAR-X status (present or absent)	Alzheimer's patients with VAR-X
Primary Exposure	Presence of VAR-X	VAR-X therapy (random assignment to treatment or placebo)
Primary Outcome	Rate of Alzheimer's progression (cognitive decline)	Change in cognitive decline (comparing the change in rate between the treatment and placebo groups)

- **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?

  ANS Randomized Control Trial would provide the strongest evidence for causality. By isolating the VAR-X intervention, if the treatment group's cognitive decline slows, the therapy is definitively proven as the cause. The prospective cohort can only show a correlation, which cannot prove that VAR-X causes the decline because other confounders might affect. A prospective cohort is good for confirming.

Terms	Prospective Cohort	Randomized Control Trial (VAR-X therapy)
Ethics	It is a **straightforward** process, mostly concerned with collecting data and consent from the patients.	This design is **more complex** than a prospective cohort, as it deals directly with clinical interventions and patient treatment.
Cost	The cost can be high over time,but surely **cheaper** than VAR-X therapy	This study is **extremely expensive** as it covers many processes, from developing the medicine until its use with patients and the analysis of the results
Time	Typically it takes a long time to observe the rate of cognitive decline over time. Also, Alzheimer's patients are often elderly, which can lead to information bias such as loss to follow-up. However, the cohort study may take **shorter time** compared to the overall process of developing a new VAR-X therapy medicine.	Comparing the entire process from the start, the experimental study will require a **longer time frame** because it includes drug development, with a significant delay before the medicine is available to patients.
Feasibility	It is **more feasible** than the experimental study since no development phase is required. The study can start by recruiting Alzheimer's patients, detecting VAR-X status, and proceeding with the study's protocol.	The experimental study is **more difficult to succeed** because it requires a drug development process, and there is a risk that the medicine will not be successful. For this reason, the study might not be feasible.

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).

  ANS The two most likely and impactful sources of bias are **information bias** and **confounding.**

Source of Bias	Detail
Information bias	• High **attrition** (loss to follow-up) due to the advanced age of Alzheimer's patients. • The possibility of **inaccurate data** when relying on interview responses from cognitively impaired participants and their caretakers. (**recall bias**)
Confounding	Confounding variables, such as age, lifestyle, smoking, and diet.

- **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).

  ANS

Source of Bias	Strategies to Minimize Biases
Information bias	• To minimize **attrition**, we could conduct follow-up visits more **frequently**. For instance, instead of the typical 2–3 year follow-up period, we could increase the frequency to every **6 months to 1 year**. • For **recall bias**, we might **validate** interview data using other health records. The interview itself should be structured with **standardized criteria** to assess cognitive decline, ensuring all participants are evaluated consistently.
Confounding	Use a matching **method** during the definition of the study population. Matching on primary confounders, such as **age**, 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?

  ANS To leverage large-scale resources for a retrospective cohort study, we must first make sure that the data fits our study's design and objectives, by following a four-step: **Collect**, **Clean**, **Transform**, and **Analyze**. First, **collecting** the raw genetic and clinical records from the biobank or EHR network. Next, **cleaning** the data to ensure accuracy by standardizing schemes and handling missing values. Then, we **transform** the raw inputs into the final structured exposure (VAR-X), outcome, and confounder variables. Finally, we can use this data to **analyze** using statistical methods.

- **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?

  ANS Traditional prospective 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 a lot of time.

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