Discussion Assignment 1

Introduction to Causal Inference

1 Instructions

Read the following two studies. Imagine you are getting ready to begin an analysis. For each study, **think** through the first three steps of the causal roadmap: (1) specifying your question, including the general system you would like to study; (2) using a SCM to represent your knowledge (or lack thereof); and (3) specifying your target causal parameter.

Please provide brief *written* answers to the questions for your group's assigned study (study 1 or 2). Please see the course website for group and study assignments. An effort has been made to create groups with a variety of backgrounds and expertise. You are encouraged to discuss with your group, but please bring your own written responses to class. **Also, be prepared to *present* your answers in class.**

1. **Step 1: State scientific question.**

   (a) What causal question or questions was this research aiming to answer?
   (b) What makes this a causal question as opposed to a purely statistical question? Give an example of a purely statistical question that the research described could address. Is the language used by the authors in describing their objective causal or statistical?
   (c) Specify the target population.

2. **Step 2: Specify a structural causal model (SCM).** Use the formal notation defined in class. Recall that W denotes baseline (pre-exposure) covariates, A the exposure or treatment, and Y the outcome. Feel free to split your covariates into $W1$, $W2$,... as needed, but please define your random variables. If needed, use Z to denote a random variable, occurring after the exposure but before the outcome.

   The data collected should give you some idea of the covariates that could be included, but there may be additional ones. When needed, we have attempted to provide additional background material to facilitate model specification. You and your group should be able to fill in the remaining material. However, do not be too concerned if you are unsure of some of the relationships in your model.

   (a) What are the endogenous variables X? What additional covariates besides those observed might be important? How would you incorporate them into the model?
   (b) Discuss your exogenous variables U. What factors might be included? Did you observe your U? Could you observe your U?
   (c) Specify your structural equations F. (Please be mindful of the notation used in class.) Do your structural equations make any assumptions about functional form?
   (d) Discuss any exclusion restrictions. In general, what are exclusion restrictions? What do they mean in words? For the study, do you feel that any are justified?
   (e) Discuss any independence assumptions. In general, what are independence assumptions? What do they mean in words? For the study, do you feel that any are justified? When might they be?
   (f) Draw at least one possible causal graph.

3. **Step 3: Define the target causal parameter.** Carefully consider the causal question or questions this research was aiming to answer.

   (a) Specify the intervention node(s).
   (b) Specify the intervention(s) of interest.
   (c) Formally express this modification of experimental conditions as an intervention on the SCM and on the causal graph.
   (d) Specify your counterfactual outcomes under the intervention(s) of interest. How are these counterfactuals defined with an SCM? What do they mean (in words)?
   (e) Using counterfactual notation, define a target causal parameter without using a marginal structural model. What does it mean (in words)?
   (f) *Bonus:* Using counterfactual notation, define a target causal parameter using a *working* marginal structural model (MSM). What does it mean (in words)?

2 Study #1: Physical activity and mortality in the elderly

Extracted and modified from: Manini TM, et. al. "Daily activity energy expenditure and mortality among older adults." JAMA. 2006 Jul 12;296(2):171-9: PMID: 16835422.

Observational studies have shown that older adults who report low physical activity levels are at elevated risk of mortality compared with those who report moderate or high levels of activity. These findings were based on questionnaire assessments of physical activity, which are subject to recall bias, are unable to account for free-living activity (energy expended during the regular course of daily living), and typically overestimate actual amounts of physical activity. Furthermore, self-reported physical activity does not provide accurate estimates of absolute amounts of activity (kilocalories per day) and thus cannot be evaluated to determine whether higher levels of activity-induced energy expenditure confers survival advantages. The most accurate and precise method of determining free-living energy expenditure uses doubly labeled water. The doubly labeled water method captures any form of physical activity ranging from purposeful exercise to simple fidgeting, whereas physical activity questionnaires generally address basic volitional activities (e.g. household chores, walking, and vigorous exercise). Although this method can address whether higher levels of free-living activity energy expenditure are associated with mortality risk, to our knowledge no studies have been conducted to determine whether free-living activity energy expenditure assessed in this way is related to longevity among older adults. The purpose of this study was to determine the association of free-living activity energy expenditure, measured using doubly labeled water coupled with resting metabolic rate, with all-cause seven year mortality in a group of high-functioning, community-dwelling older adults.

In 1997-1998, investigators from the University of Pittsburgh (Pittsburgh, Pa) and the University of Tennessee (Memphis) recruited 3075 participants aged 70 to 79 years from a random sample of white Medicare beneficiaries and all age-eligible black community residents to participate in the Health, Aging, and Body Composition (Health ABC) study. Eligibility criteria included self-reporting no difficulty in walking a distance of 0.4 km or climbing at least 10 stairs, independently performing activities of daily living, plans to live in the area for the next 3 years, and no evidence of life-threatening illnesses. An energy expenditure substudy carried out in 1998-1999 enrolled 302 individuals.

Among these individuals, free-living activity energy expenditure over a two week period was assessed using an accurate and objective measure based on doubly labeled water and indirect calorimetry. This measure captures both resting metabolic rate and total energy expenditure.

Vital status was ascertained by telephone contact every 6 months over an 8-year period (1998-2006). Date of death was verified with death certificates. Body fat was measured at the first energy expenditure visit. The following self-reported medical conditions (assessed at first visit) were confirmed by treatment and/or medication: cardiovascular disease (hypertension, coronary heart disease, myocardial infarction,

and stroke), lung disease, diabetes, hip or knee osteoarthritis, osteoporosis, cancer, and depression and were updated at the at the first energy expenditure visit. Smoking behavior was assessed during the first Health ABC annual clinic visit in 1997-1998.

Comments & Additional Questions: Think through the following additional questions and be ready to present your answers in class. (You do not need to write these up.)

1. What are some advantages/disadvantages to specifying a simplified SCM that treats self-reported medical conditions as one variable?
2. How would you incorporate specific types of physical activity on your SCM?
3. Why do you think these authors focused on total energy expenditure as the exposure of interest? For what questions might this be the most interesting intervention? For what questions might it be less interesting?

3 Study # 2: Effect of male circumcision on risk of HIV acquisition.

Extracted and modified from: Shaffer DN, et. al. "The protective effect of circumcision on HIV incidence in rural low-risk men circumcised predominantly by traditional circumcisers in Kenya: two-year follow-up of the Kericho HIV Cohort Study." J Acquir Immune Defic Syndr. 2007 Aug 1;45(4):371-9. PMID: 17558336.

Male circumcision in the prevention of heterosexual female-to-male HIV transmission has received considerable attention throughout sub-Saharan Africa. In December 2006, the US National Institutes of Health terminated early 2 randomized controlled trials (RCTs) in Kenya and Uganda after interim analyses demonstrated a marked protective effect from circumcision on HIV transmission. Now, with 3 RCTs demonstrating reductions in HIV acquisition ranging from 48% to 61%, research equipoise no longer exists with regard to the safety and efficacy of medically performed circumcision in significantly lowering the risk of men contracting HIV through heterosexual contact in areas of sub-Saharan Africa. With definitive RCT data now available, public health stake holders must consider future policy recommendations regarding circumcision alone or in combination with other prevention measures in reducing the HIV burden in sub-Saharan Africa. Although little question remains about the protective effect of male circumcision in reducing heterosexual transmission of HIV from women to men, generalizability of RCT results and experiences to populations with unique cultural and epidemiologic characteristics must be fully considered in the planning of any widespread prevention initiative. The Kericho HIV Cohort Study, located in the rural tea plantations of the southern Rift Valley Province in Kenya, offers the opportunity to study the relation between male circumcision and incident HIV in a rural population. In this study, we examined the association between circumcision and HIV infection in a cohort of adult agricultural workers and dependents after 2 years of follow-up.

The HIV and Malaria Cohort Study Among Plantation Workers and Adult Dependents in Kericho, Kenya (henceforward referred to as the Kericho HIV Cohort Study) is a 3.5-year observational, prospective cohort study aimed at estimating HIV prevalence, incidence, comorbidities, molecular epidemiology, and vaccine feasibility and acceptability among adult tea plantation workers and dependents in rural Kericho, Kenya. The Kericho HIV Cohort Study was conducted on a large tea plantation on the outskirts of Kericho, a town with a population of approximately 150,000 in the southern Rift Valley Province of Kenya. Briefly, after full review and approval by the Kenya Medical Research Institute and the Walter Reed Army Institute of Research Institutional Review Boards, 2801 adult plantation workers and dependent volunteers aged 18 to 55 years were recruited over a 6-month period beginning in June 2003. After providing written informed consent in Kiswahili or English, study volunteers were enrolled in the baseline cohort. For this 2-year follow-up sub-analysis, male circumcision study, we excluded all women (n = 1081) and HIV-positive men (n = 195) from the baseline cohort, volunteers who were lost during follow-up (n = 146), and any volunteer with incomplete data (n = 1), resulting in 1378 men for analysis.

Volunteers in the Kericho HIV Cohort Study completed extensive baseline questionnaires collecting sociodemographic, medical, behavioral, and HIV risk data. Covariates collected included religion, tribe, and self-reported sexual activity including sex with a commercial sex worker and diagnosis with a sexually transmitted infection. Circumcision status was identified at baseline with a self-report. HIV status at two years of follow-up was assessed using a standardized testing algorithm.

Comments & Additional Questions: Think through the following additional questions and be ready to present your answers in class. (You do not need to write these up.)

- When specifying your SCM, you do not need to incorporate loss to follow up, death or potential reporting bias. We will learn how to incorporate these later in the course.
- When using a marginal structural model to define the target parameter, feel free to refine your research question. However, be able to explain why the modified question is interesting.

1. How would you formally incorporate the following information in your causal model: in this population men did not alter their sexual behavior because of their circumcision status? Do you think such an assumption is warranted here? Would it be warranted if there were a program implementing adult male circumcision as an HIV prevention strategy? If you are not sure, how would you incorporate this uncertainty in your model?
2. What makes the Kericho population a potentially interesting system to study (as compared, for example, to a population enrolled in a randomized trial)? What might make it less interesting/relevant than other populations? (Think about how you are hoping to use the results of this analysis.)