Chapter 15 An Introduction to some advanced topics

Minato NAKAZAWA, Ph.D.
<minato-nakazawa@people.kobe-u.ac.jp>

As of Epidemiology (15)

Dealing with missing data

- Almost all dataset has missing: Ranging from a few for several people to many missing for a large proportion of individuals
- How to handle missing data? There are several methods, but any choice rests on assumptions about the pattern of missingness.
  - **Complete Case Method**: Exclude all individuals with (even single) missing data
    - It seems to ignore the missing data, but its justification relies on a hypothesis about why the data are missing: Missing occurs randomly, unrelated to any measured or unmeasured variables
    - By **R**, if original dataset is **X**, **subset(X, complete.cases(X))**.
    - Simple and easy, but not optimal because (1) if missing occurs not randomly, removal causes bias, (2) removal results in loss of data (Even if only 1% of individuals have missing for each variable, to use 50 variables, 39% (=1 – 0.99^{50}) of datasets are removed).
  - **Missing Indicator Method**: Instead of excluding individuals with missing data, adding a flag for each variable with missing data, typically as a new binary (0/1) variable (called as **indicator variable**), where 1 means missing. It relies on the same hypothesis (missing occurs randomly) with complete case method.
  - **Imputations** → See, next slide

Imputations

- A good tutorial paper: https://doi.org/10.1016/j.cjca.2020.11.010
- Single imputation: Filling in the missing values with one set of plausible values. Susceptible to bias if the missing values are not missing at random (non-existing data are added → less variability → narrow confidence intervals)
  - To use the mean of the nonmissing value for that variable
  - To identify important strata of individuals and use the mean of the nonmissing values for that stratum
  - To sample an imputed value randomly from the set of observed values for those whose data are not missing “hot deck imputation”
  - For longitudinal research, missing value can be imputed from that individual’s last recorded observation “last observation carried forward”
- Multiple imputation: To address the problems of bias and incorrect precision, imputing the missing value using a regression model as prediction tool.
  - To predict the most plausible value to substitute for what is missing, based on all known data, which may include study outcome
  - Complicated, but the process can be automatically applied using statistical software packages (such as **mice** and **Amelia** in R): The imputation is repeated 20 times or more (thus “multiple” imputation) and integrated (averaged across all the imputed datasets), to address the issue of false precision from adding data, with adding the error terms.
  - It appears to be a robust method that outperforms other methods in many circumstances.
  - To use **Amelia**, see https://gking.harvard.edu/amelia

Causal diagrams

- Most imputation techniques need the knowledge to clarify the relationships among variables each other → Causal diagrams are useful
- Figure 15.1 (source: Lipsky and Greenland, 2022) https://www.researchgate.net/publication/358932233_Causal_Directed_Acyclic_Graphs
  - E is associated with O if there is an open (unblocked) path between them
  - The association may be mediated via one or more other variables (M)
  - E and O are directionally connected: E is cause of O
  - If we control for M in the analysis, we block the path from E to O through M, direct effect (E → O) remains
  - E and O are independent if there is no path connecting them or every path connecting them is blocked (= directionally separated)

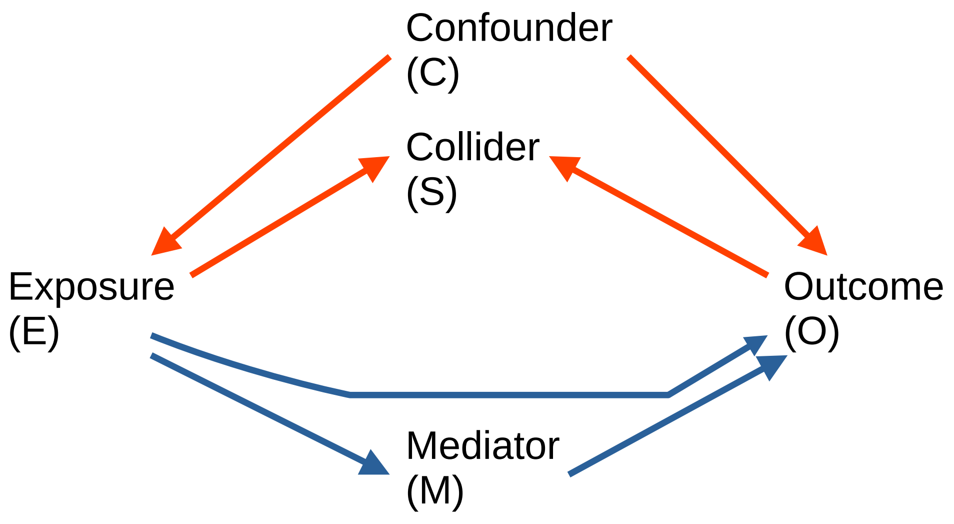

Figure 15.1 Directed acyclic graph (**DAG**) illustrating directed (or causal) and nondirected (or bias) paths.

- Nondirected paths are potential sources of bias.
  - When the third variable is associated with E and is a risk factor of O, it's confounding: E ← C → O (So called "backdoor" path, unblocked)
  - When a variable is a consequence of both E and O, it's a collider: E → S ← O, this path is blocked by the collider. If this variable is ignored, no bias, but if it's adjusted (=opening backdoor) in analyss, it causes bias.

Graphical Connection, Association, and Causation

- Collider bias explains the observations that smoking appeared to be protective against serious COVID-19: Suffering from smoking-related illness and severe COVID-19 both increase risk of hospitalization. **Those 2 risk factors are truly unassociated**.
- Assuming there are only 2 reasons for hospitalization and anyone with either risk gets hospitalized
- Table 15.1: A total of 5+45+95 people are hospitalized, no association
- Table 15.2: Wihout 855 with neither risk factor. Strong negative association exists → All negative for smoking have COVID-19, and vise versa.
- By conditioning on the **collider** of hospitalization, we may open a backdoor path between smoking and COVID-19, also called as Berksonian bias.

Table 15.1 Lack of association in general population

	Smoking	No Smoking	Total
COVID-19 infection	5	45	50
No COVID-19 infection	95	855	950
Total	100	900	1000

Table 15.2 Illustrating association among hospitalized patients

	Smoking	No Smoking	Total
COVID-19 infection	5	45	50
No COVID-19 infection	95	0	95
Total	100	45	145

Time-dependent variables

- Exposures change over time → **time-varying exposures**
  - Components of diet or chronic medication use occur daily or vary seasonally
  - Radiation exposure during a mammogram occur infrequently and sporadically
- A simplistic way is to ignore the time variation: Compare ever exposed vs never exposed → Such a dichotomous exposure definition cannot capture detailed history of exposure
- By taking the exposure history into account in the exposure definition (eg. cumulative measures of smoking such a pack-years), it's improved
- Potential confounders: unpredictably changing blood pressure, BMI, cholesterol, physical activity, exposures to sunlight, … The exposure at some time affecting the subsequent exposure (**Figure 15.2**)
- C ↔ E: C is confounder (and thus should be controlled) and causal intermediate (and thus should not be controlled) → Dilemma

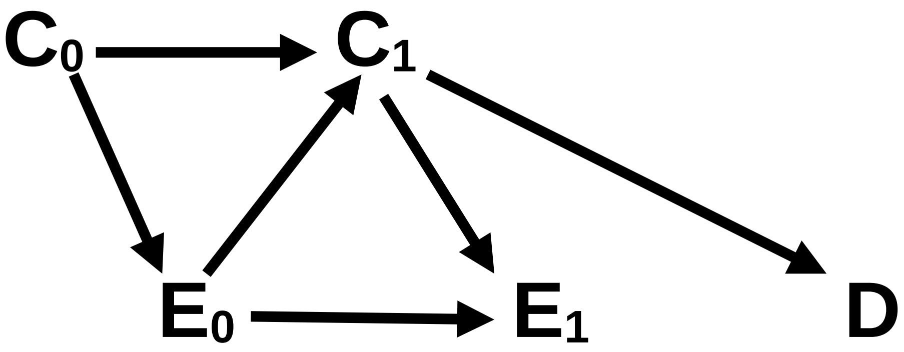

Figure 15.2 Directed acyclic graph (**DAG**) illustrating time-varying confounding.

- Consider a study evaluating the effect of asthma rescue mediation (E) on pulmonary function (D).
- The effect of medication is confounded by recent severity of asthma symptoms (C).
- Negative feedback bias the effect of treatment to result in apparent no benefit of the medication
- To solve it, g-methods (g-formula, marginal structural models, structural nested models) can be applied: These allow for stepwise feedback between time-varying treatments and time-varying confounders.

g-methods

- In Japanese, please see https://www.m.u-tokyo.ac.jp/sph/wp-content/uploads/2024/09/3ee194aec0a51beb45cd166123592ac1.pdf
- Naimi AI et al. (2017) An introduction to g methods. *Int J Epidemiol* 46(2): 756-762. PMID: 28039382; PMCID: PMC6074945. (ref.14) https://doi.org/10.1093/ije/dyw323
- g-methods estimate contrasts of potential outcomes under a less restrictive set of assumptions than standard regression methods.
  - **Inverse probability weighting** generates a pseudo-population in which exposures are independent of confounders, enabling estimation of marginal structural model parameters.
  - **g-estimation** exploits the conditional independence between the exposure and potential outcomes to estimate structural nested model parameters. In R, gesttools package can be used. https://doi.org/10.1353/obs.2022.0003
  - The **g-formula** models the joint density of the observed data to generate potential outcomes under different exposure scenarios. In R, gfoRmula (https://doi.org/10.1016/j.patter.2020.100008) package is available.

Instrumental variables

- Conditional exchangeability: The assumption of no unmeasured confounding of the exposure effect to consider that the experiences of the exposed and unexposed are exchangeable.
  - E → O (no confounding)
  - E ← C → O; E → O (blocking the backdoor path by controlling all confounders)
- This is untestable. The threat of unmeasured confounding remains one of the biggest challenges in epidemiologic research
- Using instrumental variable (Z) in Figure 15.4 can be applicable if some of the confounders remain unmeasured (U)
  - Z is associated with E, does not share any causes with O, has relation to O only through E (exclusion restriction)

- It implies
  - There is no path from Z to O that does not go through E

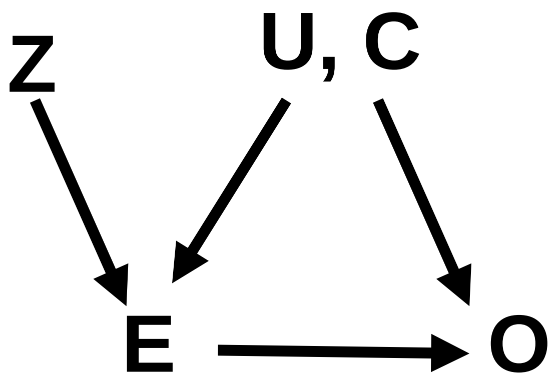

Figure 15.4 Directed acyclic graph (**DAG**) which includes an instrumental variable Z

Instrumental variable in the context of experiments

- The conditions of applicability of instrumental variable is related to random assignment to a treatment or intervention in a double-blind clinical trial.
- First, the treatment assignment is strongly associated with the actual exposure, because participants are much more likely to receive a given treatment if they are assigned to receive it.
  - (eg.) The people in treatment group and control group may change their mind about participation.
- Second, because treatment is assigned randomly, all risk factors for the outcome will be balanced between treatment groups in expectation.
  - Main goal of randomization.
- Third, because the trial is double-blind and therefore neither the patients nor the providers know the assignment, the only way in which treatment assignment can affect the outcome is through the causal action of actual treatment.
- From those 3 considerations, random assignment meets the conditions needed as instrumental variable.
- To preserve this benefit, the main effect measure of interest must be ITT (intention-to-treat), rather than actual treatment (=average treatment effect), [the effect of ITT]/[association between Z and O = measure of compliance]
  - If the compliance is perfect, denominator = 1 → the average treatment effect equals to the intention-to-treat effect.
  - If ITT effect is corrected for the amount of noncompliance to represent the effect if everyone is compliant

Instrumental variables in the context of observational studies

- The prospect for using an instrumental variable hinges on whether we can identify a variable as random treatment assignment → Quasi-experimental study
- Genetic variants that influence phenotypes offer one such possibility: Mendelian randomization studies
  - (eg.) Alcohol consumption associates with many behavioral and environmental factors. In cohort/case-control studies, those factors may confound associations between alcohol use and health outcomes. But flusher who has malfunctioning variant gene in ALDH feel unpleasant to drink alcohol and thus associates with less alcohol consumption. Having less active ALDH may be an instrumental variable to study the health effect of alcohol consumption, but there are many challenges.
    - Can we find genes affecting the exposures of interest?
    - Are the associations between genotype and phenotype too weak to be helpful?
    - Can we be certain that the other assumptions for instrumental variables are met?
- Beyond Mendelian randomization, there are several other sources of random variation which can be exploited in instrumental variable analysis → “preference”
  - Physicians may have a preference to prescribe one medication over another
  - Some hospitals may prefer more aggressive treatment of certain conditions
  - Such provider, facility, or even region-based preferences can be used as an instrumental variable.

Challenges with instrumental variable analyses

- The biggest challenge is that we can never prove that a variable is a valid instrument
  - The first condition (association between the instrument and the exposure) can be empirically verified to some extent. No clear cut-off
  - The second and third conditions may be theoretically justified, but not empirically proven.
    - The balance of risk factors may indicate that measured risk factors for outcome are imbalanced by instrument status
    - But if balanced, unmeasured risk factors could still be unbalanced
    - If co-interventions affect the outcome directly, the absence of such co-interventions cannot be proven.
  - Even minor violations of the instrumental variable conditions may result in large biases of unpredictable direction.
- Second challenge relates to the fact that, even if several variables are controlled as instrumental variables, other conditions will affect the correct interpretation: It cannot be empirically verified.
- If a valid instrumental variable is defined, implementation is straightforward through 2 stages
  - Predicting E as a function of Z and potentially measured C
  - Predicting O as a function of predicted E and the same set of C

Quantitative bias analysis

- Anyway, bias (selection bias, confounding, misclassification) remains in almost all epidemiologic studies.
  - (eg.) Misclassification of exposure about medication (false positive and false negative)
  - (eg.2) Misclassification of disease outcome (overdiagnosis occurs more in exposed)
  - (eg.3) Misclassification in confounding (some confounders are imperfectly measured or not recorded at all)
- Common approach for bias: Discussing the biases are small
- Alternative, more powerful approach is Quantitative bias analysis (Table 15.3)
  - Simple sensitivity analysis: one fixed value assigned, one bias is analyzed, single revised estimate of association is given, random error is not fully incorporated
  - Multidimensional sensitivity analysis: 2 or more values assigned, one bias is analyzed, 2 or more revised estimates are given, random error is not fully incorporated
  - Probabilistic analysis: probability distribution assigned, one bias is analyzed, frequency distribution of revised estimates is given, random error is fully incorporated
  - Multiple bias modeling: probability distribution**s** assigned, multiple biases are analyzed, frequency distributions of revised estimates are given, random error is fully incorporated

Simple and multidimensional sensitivity analyses

- Using expected impact of the systematic error using bias parameters
- A simple sensitivity analysis assesses the impact on the study findings of assuming one alternative fixed value for the bias parameters.
  - In **Table 15.4**, assuming that our measure of exposure has a specificity of 80% and 100% sensitivity, observed number of unexposed total is 800*0.8=640; exposed total is 200+(800-640)=360; Among 150 unexposed cases, 150*(1-0.8)=30 are misclassified as exposed and thus exposed cases are 100+30=130 (actually among (800-640), 30 develop disease); unexposed cases become 150-30=120
  - Observed RR = 0.36/0.19 = 1.9, which is much smaller than true RR 2.7, due to nondifferential misclassification
  - In real world, we don't know truth. If we start from observed data and assume 80% specificity and 100% sensitivity, 640/0.8=800 as true total unexposed, 800*(120/640)=150 as exposed cases, 360-(800-640)=200 as total exposed, 130-(150-120)=100 as exposed cases (**back-calculation**)

Table 15.4 Exposure misclassification

	Truth		Observed	
	Exposed	Unexposed	Exposed	Unexposed
Diseased	100	150	130	120
Total	200	800	360	640
Risk	0.5	0.19	0.36	0.19
Risk Ratio	0.5/0.19 = 2.7		0.36/0.19 = 1.9	

- As shown in Table 15.5, various sets of sensitivity and specificity can be assumed, and accordingly, various "truth" can be calculated, which enable to estimated various "true" RR.
- Assuming the range of specificity as 0.7-0.9 and the range of sensitivity as 0.85-1.0, possible RRs range from 2.2 to 5.4.
- The range of those RRs is much wider than the conventional confidence intervals (in this case, 95% CI of RR 2.7 is 2.2-3.3)

Probabilistic and multiple bias modeling

- Multidimensional sensitivity analysis does not incorporate any outside information or prior view about which estimates are most plausible
- Probabilistic analysis considers probability distributions (uniform, triangular, trapezoidal, …) for the bias parameters rather than the sets of plausible values.
- Values for the bias parameters are repeatedly (1000 times or more) drawn from prespecified distribution. The results are accumulated to generate the frequency distribution of the results. Percentiles can be reported. Finally random errors are incorporated.
- In R, **episensr** package enable it.
  https://cran.r-project.org/web/packages/episensr/vignettes/episensr.html
  https://doi.org/10.1093/ije/dyad053 (Explanation for R and SAS)
- Table 15.6 provides the example of probabilistic bias modeling
- By simultaneously considering multiple sources of bias, multiple bias modeling is also possible.
- Critics of bias modeling is the subjective (or educated guesses) choice of bias parameter distributions, but it's important in causal inference based on weak associations but needed in public policy or clinical practice.
- E-value is suggested as quantitative measure of unmeasured bias
  https://doi.org/10.1016/j.jclinepi.2023.09.014
  https://doi.org/10.1093/ije/dyaa127
  https://www.igaku-shoin.co.jp/application/files/8516/5208/5051/114.pdf (Japanese)

FURTHER READING

- Baraldi AN, Enders CK. An introduction to modern missing data analyses. J Sch Psychol. 2010 Feb;48(1): 5-37. PMID: 20006986. https://doi.org/10.1016/j.jsp.2009.10.001
- Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol. 2015 Apr 7;15:30. PMID: 25880850; PMCID: PMC4396150. https://doi.org/10.1186/s12874-015-0022-1
- Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393. PMID: 19564179; PMCID: PMC2714692. https://doi.org/10.1136/bmj.b2393
- Imai, K. (2021) Causal Directed Acyclic Graphs. Harvard Univ. https://imai.fas.harvard.edu/teaching/files/DAG.pdf
  - Freesoft: https://dagitty.net/ (Tutorial: https://dagitty.net/learn/index.html)
- (For Japanese Students, the books below are very good resource)
  - 佐藤俊哉 (2024)『宇宙怪人しまりす⬚統計よりも大事なことを学ぶ』朝倉書店
  - 林岳彦 (2024)『はじめての統計的因果推論』岩波書店