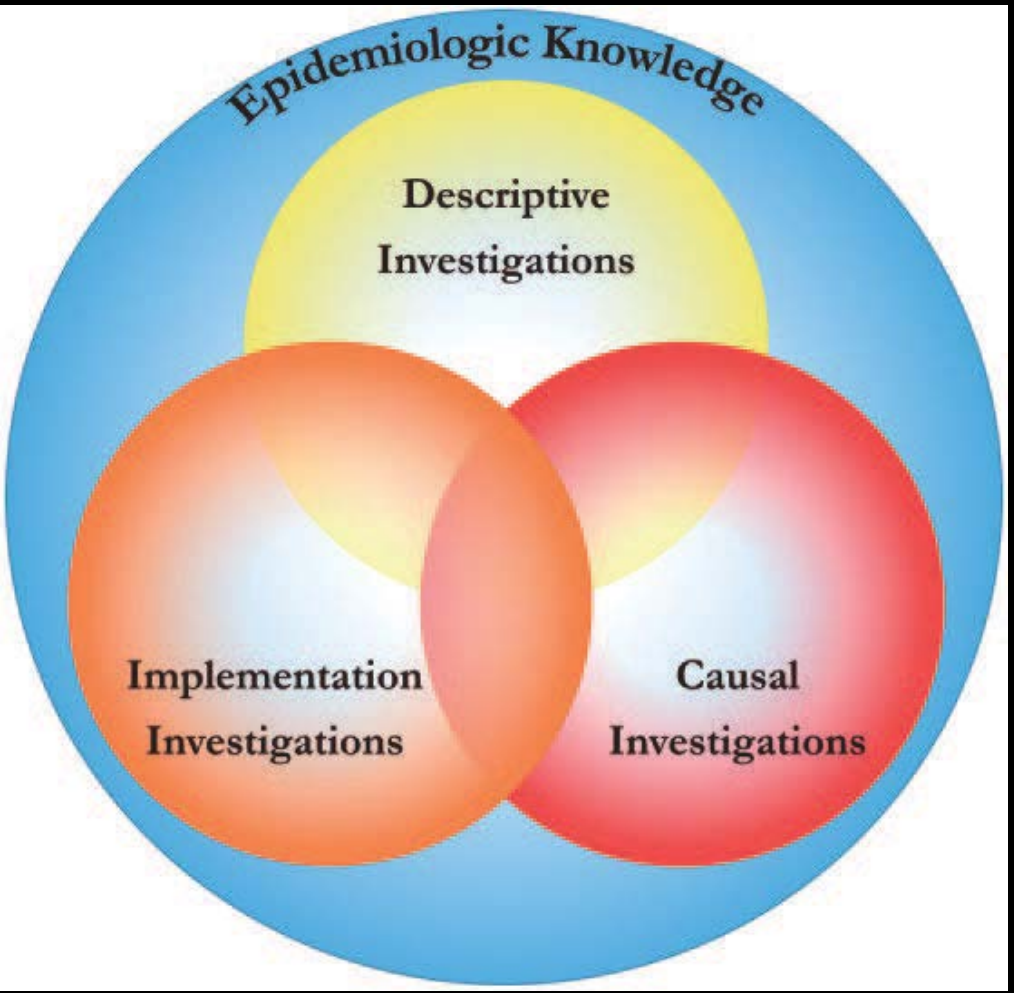
Overview of issues surrounding the role of representative sampling in epidemiology

Hormuzd A. Katki, Ph.D.

Senior Investigator
Division of Cancer Epidemiology and Genetics

3 roles of epidemiology



- **Both Surveillance and Policy/Action require**
  - **Population/subgroup-representative absolute rates**
- **Causal science doesn't necessarily require absolute rates**
  - **Multiplicative relative rates**
  - **Causal inference is toward counterfactual populations, not real ones**
  - **Art of science is finding the best non-representative population to easily discover a "law of nature"**

Lau, et al, *Int J Epidemiol*, 2018

Population representativeness is expensive. Lower priority if biospecimens are needed.

- The NHANES survey recruits ~8,000 people per cycle at cost of $100M: ~$10,000 per recruit
  - ~70% response rate in African-Americans but <50% in Asian-Americans
- Non-probability samples cost ~$1,000 per recruit in the US
  - High cost in US is why many epidemiologic studies are done in countries with lower costs but good medical infrastructure
  - <10% response rate
  - Example: UK Biobank
    - 500,000 volunteers recruited through clinics
    - 5.5% response rate
    - Volunteers have half the mortality rate of the UK population
    - Gold mine of biospecimens: Blood, urine, complete physical exam, linkage to health records, imaging subset, longitudinal subset

Epidemiologic study designs must be both statistically efficient and cost efficient

- If relative rates and counterfactuals suffice
  - Case-control studies are statistically efficient and cheap
  - IF: unbiased retrospective exposure assessment is possible in cases
    - Not typically true for biomarkers, which may well be affected by disease or by disease treatment
    - "Internal validity"

- If retrospective exposure assessment not possible, need prospective biospecimens
  - Means case-control studies are not useful
  - Must have a prospective cohort with biospecimens banked away at baseline

Why be population representative, if all you need are relative risks?

- **Representativeness protects against unmeasured *effect modifiers***
  - **True Model:**

$$E(Y|A,U) = \beta_0 + \beta_1 A + \beta_2 U + \beta_3 A \times U. \quad (\text{assume } A \perp\!\!\!\perp U)$$

  - **If U is unavailable, we are implicitly marginalizing over U:**

$$E(Y|A) = (\beta_0 + \mu_U \beta_2) + (\beta_1 + \mu_U \beta_3) \times A$$

  - **$(\beta_1 + \mu_U \beta_3)$ is the correct marginalized effect of A when U is unavailable**

- **When do we have an unbiased estimate of $(\beta_1 + \mu_U \beta_3)$?**
  - **Not under non-representative sampling, because the implicit μ_U in the sample will not be the correct μ_U**
    - **Unless $\beta_3=0$, which means no effect modification**
    - **Problem persists if the effect of A is misspecified (e.g. A should be quadratic) and A and U are dependent, leading to an A × U interaction in the marginalized model**
  - **Representative sampling naturally marginalizes correctly**

Adapted from Elliott, *JRSS-A*, 2016

Compromise: Improve external validity at the analysis stage

- Use a representative survey as a frame to develop ”pseudoweights” for the epidemiologic study
  - Long history in web surveys, where the sampling frame is typically unknown
- Develop a model for propensity to be the survey versus the epidemiologic study
  - Develop pseudoweights based on
    - Inverse of the propensity score
    - Subclassification on the propensity score
- Other approaches possible, will not be discussed
  - Imputation of variables into the survey using epidemiologic data

Rivers, *JSM Proc*, 2007 ; Lee and Valliant, *Sociol Methods Res*, 2009
Elliott, *Survey Prac*, 2013 ; Elliott et al, *Stat Sci*, 2016; Wang et al, submitted

Optimizing surveys and epidemiologic studies for reducing bias in epidemiologic studies

- Harmonizing variables in surveys & epidemiologic studies
  - Problems of splitting vs. lumping categories
  - Ask questions in the exact same ways
    - Hopefully optimally determined by psychologic research
- Measure all key sociologic variables that might determine participation in epidemiologic studies
  - Epidemiologic studies that collect biospecimens typically do not ask many sociologic questions
    - Especially on socio-economic status
- Hopefully surveys oversample people who are undersampled in epidemiologic studies
  - Otherwise survey is of minimal value for reducing bias

How can we identify the realistic population base that the analysis is weighting up to?

- Nothing can replace formal probability sampling
  - Epidemiologic studies don't merely undersample subgroups, but do not sample many subgroups
    - E.g. ~7% of Americans do not have access to a phone
  - It is hard to believe we could ever weight up a non-probability sample, even with modeling assumptions, to literally represent the US and all important subgroups

- What population base is an analysis weighting up to?
  - Can we look at the provided demographic information
  - Examine the propensity model in a probability sample, see which people have the lowest probabilities of participation
    - Examine their demographics
    - Calculate the total number of "basically non-sampled" people in the population
  - Which subgroups does our analysis represent, and which not?

Sensitivity analyses that assess the amount of bias not removed by the analysis

- The proposed analyses can reduce bias, but never eliminate it
- Can we look at outcomes available in both the probability and non-probability samples
  - How much bias can the proposed weights correct for outcomes that are observed in the probability sample?
  - Might be provide an informal assessment of how much bias can be corrected

We need more realistic propensity models than generalized linear models

- Propensity models aim to model human biases and whims that lead to participation or not
  - These biases and whims are surely non-linear and complex
- Other possible propensity models
  - Machine learning algorithms, such as regression trees
    - Bayesian Additive Regression Trees (BART) (Xu, Daniels, Winterstein Biometrics 2018)
    - BART outperformed logistic regression propensity scores when there are many covariates with interactions or non-linearity
- Sensitivity to propensity model misspecification
  - Surely the propensity model is misspecified

Linking surveys and epidemiologic studies to ”big data” sources to gather variables that might predict participation

- Survey and epidemiologic questionnaires must be limited to reduce participant fatigue

- But other linkable ”big data” sources from “data brokers” may have information that might predict participation
  - Akin to survey paradata
  - Residence history
  - Data on your purchases
  - LexisNexis “health risk scores”
  - Databases of online habits
    - Social media
    - TV habits

- This information needs to be used ethically

Conclusions

- Representative epidemiologic studies are expensive
  - If biospecimen collection is critical, often not worth the sample cut required to stay in budget
  - We have no choice but to improve external validity with new analytic techniques

- Methods discussed today can use survey data to improve external validity of epidemiologic analysis, but have a long way to go
  - Measure and harmonize all key sociologic variables predicting participation in epidemiologic studies
    - May require linkage “big data” databases
  - The new methods cannot replace representative surveys
    - We need to better understand what populations that these new methods can weight up to and cannot weight up to
    - Understand better much bias cannot be removed by these methods
    - Propensity models for participation need to more realistic and are surely misspecified