Potential Collaborations between Topic Group 5 (Design) and other STRATOS Topic Groups

Mitchell H. Gail

Biostatistics Branch, Division of Cancer Epidemiology and Genetics

NIH NATIONAL CANCER INSTITUTE

June 3, 2019

Topic Group 5 (Design)

- Aim: Provide accessible and accurate guidance on the design of observational studies
- Members: Suzanne Cadarette, Mitch Gail (co-chairs), Gary Collins, Stephen Evans, Neil Pearce, Peggy Sekula, Neus Valveny, Elizabeth Williamson, Mark Woodward
- Expository Level 1 paper on “Design Choices for Observational Studies of the Effect of Exposure on Disease Incidence” in review
- Potential papers: role of practical constraints, ameliorating potential biases, generalizability, comparative effectiveness studies
- Potential for collaboration on design issues with other Topic Groups

Missingness (TG1,TG8)

- Missing not-by-design in cross-sectional samples (survey non-response)
  - E.g. Controls or cases who participate in case-control study versus those who refuse
  - Renewed interest in survey community for inference from non-probability samples (Elliot and Valliant, Statistical Science, 2017)
- Using non-probability-based data: Pseudo-randomization
  - P(X) is proportion with variables X in a probability sample (or census) from the target population (reference).
  - Q(X) is the proportion with X in the non-probability sample.
  - Weights proportional to P(X)/Q(X) used to reweight the non-probability sample for inference on the variable of interest, Z.
  - Design issues: How to choose X; sample size implications of weighting and estimating weights on e.g. case-control analyses; additional studies to test if missingness does not depend on the variable of interest, Z

Missingness (TG1,TG8)

- Using non-probability-based data: Superpopulation model approach (imputation)
  - Want to estimate proportion exposed in entire population, P(Z)
  - $\hat{P}(Z \mid \text{responders})$ known; $\hat{P}(Z \mid \text{non-responders})$ not available
  - We have estimate of non-responder X distribution $\hat{P}(X \mid \text{non-responder})$
  - Assume P(Z|X, responders)=P(Z|X, non-responders)=P(Z|X).
  - Then

$$\hat{P}(Z) = \pi \hat{P}(Z \mid \text{responders}) + (1-\pi)\sum_{x} \hat{P}(Z \mid X, \text{responders})\hat{P}(X \mid \text{non-responders})$$

π = probability of responding among the original random sample

  - Design issues: Choosing X; Special studies to validate conditional independence assumption; Special studies to compare validity of the two approaches

Missingness by Design (TG1,TG8)

- 2-Phase sampling from a cohort (case-cohort; nested case-control)
  - Ancillary variables available on all cohort members used for improved imputation or survey calibration in second phase
    - Survey calibration/raking (Breslow et al. 2009 AJE case-cohort; Stoer and Samulesen, 2012, LDA)
    - Multiple imputation (Keogh 2018, Handbook of Statistical Methods for Case-Control Studies)
  - Augmenting or stratifying the designs
- Outcome-dependent sampling
  - Population-based case-control study
  - Longitudinal binary series
    - Oversample those with more internal variation; those with at least 1 binary outcome

      Schildcraut et al 2018 Epidemiology
  - Oversample long survival times (J Yu et al, 2016 J Stat Plan Inf)
  - Family studies with >k affected family members
  - Multistate models with oversampling of subjects prevalent in some states

Minimizing/Assessing NMAR-Missingness

- Cohort studies
  - Outcome data completeness
    - Active surveillance system; improved record linkages
    - Follow-on studies to assess completeness of exposure-specific outcome ascertainment
  - Exposure and covariate completeness (e.g. item non-response)
    - Questionnaire design; interactive data acquisition systems
- Case-control studies
  - Selective participation cases and controls
    - Studies on how to improve participation rates
    - Follow-on studies to assess representativeness of cases and controls
  - Missing exposure and covariate data
    - Questionnaire design; interactive data acquisition systems
    - Follow-on studies to assess MAR given disease status

Dose-Response (TG2, TG3)

The dose-response curve, $E(Y|X)=h(X)$

- Global features
  - Slope — Large var(X)
  - Curvature, plateau — Design to detect deviations from the linear trend
- Local features
  - Low-dose response — Many observations near X_0
  - Good local fit everywhere — Observations throughout the range of X

- Possible adaptive designs to refine initial estimates of h(X)
- Multivariate X

Reducing Measurement Error (TG4,TG6)

- Improve and pilot questionnaires or assays before main study
  - Evaluate current questionnaires or assays for bias, CV, intraclass correlation
  - Devise better questionnaires or assays
    - RR of cervical cancer and HPV16/18: 4-8 with *in situ* hybridization; 180 with DNA/PCR
  - Challenges with monitoring devices
- Use right design
  - Cohort or sub-sampling cohort to reduce recall bias and reverse causation bias
  - Outcome assessment independent of analyte
  - Exposure assessment independent of case-control status
  - Balance cases and controls within batches to reduce bias from batch effects
  - Increase sample size to compensate for non-differential error
- Real-time quality control of lab assays: Replication of standard samples within and across analysis batches

Validation and Replication To Reduce Impact of Measurement Error (TG4)

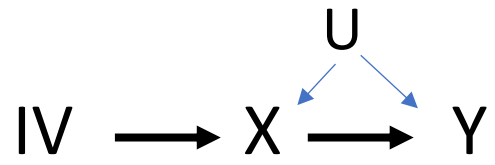
- Validation subsampling against “Gold Standard” data
  - Example: Time-to-event and covariates measured with error in electronic health records.
    - Random sample for chart-review validation. Raking calibration (Shaw, Oh, Lumley)
    - More informative missing-at-random validation designs
  - Detect and/or correct for differential error
    - Exposure measurement error in cases and controls
    - Outcome measurement error in exposed and non-exposed
- Replications within individual: at each time; across time
  - To get more precise exposure estimate

Assay Calibration in Multicenter Studies (T4)

- Example: Vit D assays for nested case-control studies of colon cancer in 21 cohorts worldwide (Gail et al, SIM 2016). Different assays used in various studies.
- Reference laboratory values regressed on local measurements to calibrate local values
- Design issues for further investigation
  - How large should the calibration samples be?
  - Should they be random samples or spread out across the local lab values?

Mendelian Randomization (TG7)

- Sample size, MSE considerations
- Meta-analyses with various instruments (Burgess SMMR 2016)
- Subsampling to get exposure data (Pierce, Burgess AJE 2013)
- Testing key assumptions
  - 1. The IV is associated with the risk factor X
  - 2. The IV is not associated with any confounder U
  - 3. The IV is conditionally independent of the outcome Y given X
- Are assumptions 2 and 3 valid for genetic risk scores based on many genetic variants? (Qi, Chatterjee Na Comm 2019)



- Pragmatic trials with randomized assignment the instrument

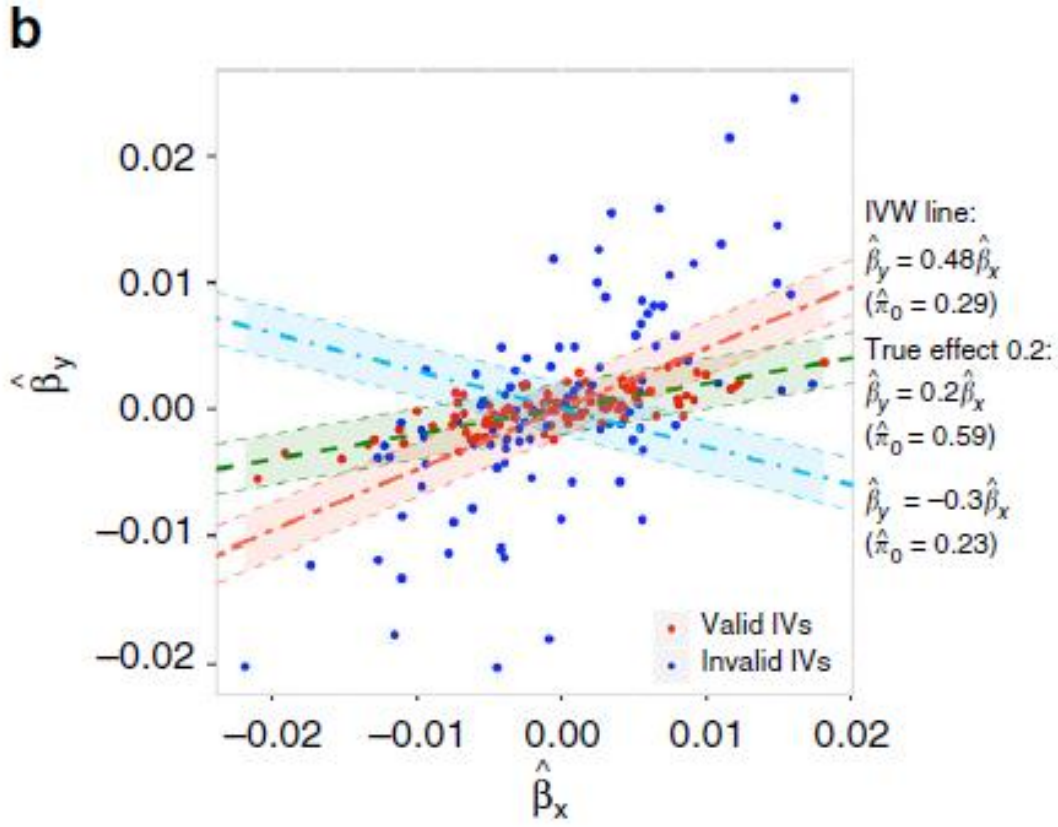
Plot of regression of outcome Y on SNP versus regression of exposure X on SNP (Qi Nat Comm 2019)

If $\beta_{gY} = \beta_{gX}\beta_{XY}$, then

$(\hat{\beta}_{gX}, \hat{\beta}_{gY})$ should lie on line

with slope β_{XY}

Red SNPs satisfy IV assumption;
Blue SNPs do not



Causal Inference (TG7)

Experiments or other data to test features of directed acyclic graphs

- Randomize the values in the DAG : randomize V1 for V1 $\longrightarrow$ V2
- Instruments IV $\longrightarrow$ V1 $\longrightarrow$ V2 vs IV $\longrightarrow$ V1 $\longleftarrow$ V2

- Studies with the same person given both treatments at different times. If no order effect or carry-over effect, we get to observe (Y_0,Y_1).
  - Compared to individually randomized trial with one outcome observed on each person
    - Validity of the no order effect/no carry-over assumptions
    - Efficiency if assumptions are satisfied
  - In observational data
    - No order effect/no carry-over assumptions vs no unmeasured confounder assumptions
    - Confounding by indication better controlled by within person design, absent order and carry-over effects
  - Design issues: scope for more use of within-person designs?; tests of required assumptions

Preparations for High-Dimensional Data Analysis (TG9)

- Standardization for obtaining and handling samples (e.g. microbiome)
  - Sample collection methods
  - Storage conditions and stability of samples/analytes in storage
  - Analyte extraction procedures
- Pilot evaluation of multiplex assays
  - Laboratory conditions; optimization for only some analytes?
  - Preprocessing electronic assay signals
  - Reliability studies to eliminate bad assays
    - Intraclass correlation (across batches or times within individual); coefficient of variation
- Understanding the study population
  - Sampling from the target population versus samples of convenience
  - Measurement/control for confounders (e.g. population stratification)
- Real-time assay quality control
  - Standard samples to identify batch effects
  - Balancing e.g. cases and controls within batch
- Data management system for reproducible research

Power Calculations and Simulations for High-Dimensional Data

- Simulating joint distribution of risk factors: analytic models versus resampling from real data
- Modeling the effects of risk factors on outcome: analytic models, mixture models from real data
- Impact of simulation methods on planning for various types of inference
  - Unsupervised clustering
  - Feature discovery
  - Risk prediction
  - Structure discovery: networks, pathways

Analytic Issues in High-Dimensional Data Analysis (TG9)

- Risk factor discovery (e.g. GWAS, microbiome)
  - Outlier detection of technical failures and/or robust statistics
  - Statistical control of false discovery
  - Independent statistical validation study of discovered features (criteria and design)
  - Follow-up studies to elucidate mechanism
- Validation of inferred pathways or networks
  - Independent validation of correlation structures?
  - Experiments?
- Risk prediction based on high-dimensional data
  - Model development(variable selection/filtering, tuning)
  - Validation of calibration in several independent cohorts
    - Selecting validation cohorts (For what target population? Required number of events?)
    - Missing predictor data in validation samples; supplemental sampling?
  - Validation of discriminatory accuracy

Summary

- New scientific challenges, new technologies, and new analytic methods may induce innovations in design
- Keep **design** in mind when devising new methods or offering guidance in your topic groups
- Think of TG5 for potential collaborations on design