Combining Data Sources to Produce Nationally Representative Estimates of Hospital Encounter Characteristics

Jay Breidt

NORC at the University of Chicago

2023 FCSM Research and Policy Conference

Hyattsville, Maryland

Joint work with Dean Resnick, NORC; Geoffrey Jackson and Donielle White, NCHS

Background, I

- National Hospital Care Survey (NHCS)
    - conducted by the National Center for Health Statistics (NCHS)
    - stratified simple random sample of non-federal and noninstitutional hospitals with six or more staffed inpatient beds
    - subject to hospital-level nonresponse
    - responding hospitals, A, provide (essentially) complete **encounter data** for all patients for 2020
- Invaluable research resource for patterns of health care delivery and utilization in the United States
    - patient demographics, diagnoses/procedures, length of stay
    - linkable to external data sources including National Death Index and Centers for Medicare & Medicaid Services data
    - 2020 data includes critically important hospital records for the first year of COVID patients

Background, II

- Focus here on A = responding NHCS inpatient hospitals
  - provide (essentially) complete encounter data for all patients for 2020
- Combine A with other data sources in order to produce nationally representative estimates
- Proprietary commercial database, B
  - nonprobability "sample" of participating hospitals
  - good coverage of hospital population by various measures (e.g., geographic dispersion)
  - participating hospitals provide (essentially) complete encounter data for 2020
- Hospital population info from Healthcare Cost and Utilization Project National Inpatient Sample (HCUP-NIS), C
  - essentially a census of hospitals, though a sample of patients within hospitals

Challenges and opportunities

- **Massive data** at the encounter level
  - number of hospitals in $A \cup B$ is **hundreds** out of **thousands** of US hospitals
  - number of encounters in $A \cup B$ is **tens of millions**
- **No linkages!**
  - deidentified hospitals in B
  - data use agreement **precludes** linking hospitals in A and B
  - no hospital identifiers in C, hence cannot link to A or B

Estimation goals: produce nationally-representative estimates

Goal	Data	Controls	Weights	Release
1	Both $A \cup B$	national summaries (using HCUP-NIS)	hospital level	national estimates
2.1	Only A	Goal 1 estimates (+ nat'l summaries)	encounter level	restricted-use data file
2.2	Subsample of A	Goal 1 estimates (+ nat'l summaries)	encounter level	public-use data file

Goal 1: Combined, weighted dataset for national estimates

- A is a probability sample
  - known inclusion probabilities, but not all hospitals respond
  - potential **differential nonresponse**
  - use information available for respondents and population to model propensity to respond
- B is not a probability sample
  - potential **differential participation**
  - use information available for participants and population to model propensity to participate
- **Caution:** no way to know how response and participation propensities might interact
- Use modeled propensities to construct hospital-level weights

Modeling hospital response propensities for NHCS

- $S \subset U$ is stratified NHCS sample, with known inclusion probabilities $\pi_h > 0$
  - stratification is determined by bed size, type of hospital, and rural/urban designation
- Define $A_h = 1$ if hospital h responds to NHCS and $A_h = 0$ otherwise and define $A = \{h \in S \subset U : A_h = 1\}$
- **Pseudo-log-likelihood criterion** is

$$\sum_{h \in U} \frac{\mathbf{1}_{\{h \in S\}}}{\pi_h} A_h \log\left(\frac{\rho_h}{1-\rho_h}\right) + \sum_{h \in C} \log\left(1 - \rho_h\right)$$

- Assume logistic model for ρ_h
- Because covariates are entirely categorical, can fit using standard logistic regression software

Modeling hospital participation propensities for proprietary

- Define $B_h = 1$ if $h \in U$ participates in proprietary and $B_h = 0$ otherwise and define $B = \{h \in U : B_h = 1\}$
- Define the **participation propensity**, $\gamma_h = \mathrm{P}\left[B_h = 1\right]$ and assume a logistic model
- The log-likelihood for estimation of parameters in γ_h is

$$\sum_{h \in B} \log\left(\frac{\gamma_h}{1-\gamma_h}\right) + \sum_{h \in C} \log\left(1-\gamma_h\right).$$

- Since $\boldsymbol{x}_h$ is entirely categorical, we can again use standard logistic regression software to maximize the log-likelihood

Goal 1 hospital-level weights, I

- Once both propensity models are fitted, we construct hospital-level weights:

$$w_h^A = \frac{1}{\pi_h \widehat{\rho}_h}, h \in A; \quad w_h^B = \frac{1}{\widehat{\gamma}_h}, h \in B$$

  - (constant within cells because covariates are categorical)
- We combine the data with a **separate dual-frame estimator**, by first choosing $\lambda \in (0,1)$ and then computing national estimates as

$$\sum_{h \in A \cup B} w_h^{AB} \sum_{i \in H_h} y_{hi} = \lambda \sum_{h \in A} w_h^A \sum_{i \in H_h} y_{hi} + (1-\lambda) \sum_{h \in B} w_h^B \sum_{i \in H_h} y_{hi}$$

  where y_{hi} is a measurement for encounter record i in hospital h and H_h is the entire set of encounter records
  - we choose $\lambda = n_A/(n_A + n_B)$

Variance estimation for Goal 1

- Variance of the separate estimator:

$$\lambda^2 \text{Var}\left(\widehat{T}_A\right) + (1-\lambda)^2 \text{Var}\left(\widehat{T}_B\right) + 2\lambda(1-\lambda)\text{Cov}\left(\widehat{T}_A, \widehat{T}_B\right),$$

  where sampling covariance term cannot be determined
  - best case: NHCS respondents are unlikely to be participants, and vice-versa
  - worst case: NHCS respondents and participants are likely to be the same
- **Stratified delete-a-group jackknife** variance estimator, with B serving as its own stratum
  - treats A and B as independent: $\lambda^2 \widehat{V}_A + (1-\lambda)^2 \widehat{V}_B$
  - accounts for uncertainty due to estimation of propensity models

Goal 2 encounter-level weights, I

- Given the combined national estimates, need to reweight **only the NHCS data** to construct
  - Goal 2.1: weighted restricted-use data set that reproduces key national estimates
  - Goal 2.2: subsample of Goal 2.1 data set to be released as public use file
- **Key considerations:**
  - no proprietary microdata will be released
  - proprietary data only appear in the national estimates that are used as controls for the new weights
  - achieving controls requires weights that vary across encounter records within hospitals
  - try to minimize variation of weights within hospitals

Goal 2 encounter-level weights, II

- Vector of key national estimates,

$$\widetilde{T}_{\boldsymbol{z}} = \sum_{h \in A \cup B} \sum_{i \in H_h} w_h^{AB} \boldsymbol{z}_{hi}$$

  - $\boldsymbol{z}_{hi}$ includes coarsened diagnosis codes, discharge status, length of stay, age group, sex, newborn status

- Goal 2 is to find encounter-level weights $\{w_{hi}^{A}\}_{h \in A}$ that vary as little as possible within hospitals while satisfying

$$\widetilde{T}_{\boldsymbol{z}} = \sum_{h \in A \cup B} \sum_{i \in H_h} w_h^{AB} \boldsymbol{z}_{hi} = \sum_{h \in A} \sum_{i \in H_h} w_{hi}^{A} \boldsymbol{z}_{hi}$$

- This is a (large) survey calibration problem

Goal 2 encounter-level weights, III

- Generalized regression (GREG) version of this calibration is obtained via

$$T_y^* = \sum_{h \in A} w_h^A \sum_{i \in H_h} \left(y_{hi} - \boldsymbol{z}_{hi}^\top \widehat{\boldsymbol{\beta}}_N \right) + \widetilde{T}_{\boldsymbol{z}}^\top \widehat{\boldsymbol{\beta}}_N$$

where

$$\widehat{\boldsymbol{\beta}}_N = \left(\sum_{h \in A} \sum_{i \in H_h} w_h^A \boldsymbol{z}_{hi} \boldsymbol{z}_{hi}^\top \right)^{-1} \sum_{h \in A} \sum_{i \in H_h} w_h^A \boldsymbol{z}_{hi} y_{hi}$$

Goal 2 encounter-level weights, IV

- GREG version of these weights is (for $h \in A$)

$$w_{hi}^A = w_h^A \left\{ 1 + \left(\widetilde{T}_{\boldsymbol{z}} - \widehat{T}_{A\boldsymbol{z}} \right)^\top \left(\sum_{h \in A} \sum_{i \in H_h} w_h^A \boldsymbol{z}_{hi} \boldsymbol{z}_{hi}^\top \right)^{-1} \boldsymbol{z}_{hi} \right\}$$

- The GREG weights are calibrated to the combined estimates:

$$T_{\boldsymbol{z}}^* = \sum_{h \in A} \sum_{i \in H_h} w_{hi}^A \boldsymbol{z}_{hi}^\top = \widetilde{T}_{\boldsymbol{z}}$$

- We use a closely-related raking approach

Goal 2 variance estimation, I

- GREG can be written

$$\begin{aligned} T_y^* &= \widehat{T}_{Ay} + \left(\widetilde{T}_{\boldsymbol{z}} - \widehat{T}_{A\boldsymbol{z}}\right)^\top \boldsymbol{\beta}_N + \left(\widetilde{T}_{\boldsymbol{z}} - \widehat{T}_{A\boldsymbol{z}}\right)^\top \left(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N\right) \\ &\simeq \left\{\widehat{T}_{Ay} - (1-\lambda)\,\widehat{T}_{A\boldsymbol{z}}^\top \boldsymbol{\beta}_N\right\} + (1-\lambda)\,\widehat{T}_{B\boldsymbol{z}}^\top \boldsymbol{\beta}_N \end{aligned}$$

  - $\lambda = 1$: estimator ignores proprietary sample B
  - $\lambda = 0$: like ordinary GREG, with model-based predictions for B instead of U
  - $\lambda \in (0, 1)$: uses a shrunken version of the model-based predictions
- Variance estimation: first term is A-only, second term is B-only

Goal 2 variance estimation, II

- Focusing on "sampling" error only,

$$\begin{aligned}\mathrm{Var}\left(T_y^*\right) &\simeq \mathrm{Var}\left(\widehat{T}_{A,y-(1-\lambda)\boldsymbol{z}^\top\boldsymbol{\beta}_N}\right) \\ &\quad +(1-\lambda)^2\boldsymbol{\beta}_N^\top\mathrm{Var}\left(\widehat{T}_{B\boldsymbol{z}}\right)\boldsymbol{\beta}_N \\ &\quad +2(1-\lambda)\mathrm{Cov}\left(\widehat{T}_{A,y-(1-\lambda)\boldsymbol{z}^\top\boldsymbol{\beta}_N}, \widehat{T}_{B\boldsymbol{z}}^\top\right)\boldsymbol{\beta}_N\end{aligned}$$

  - covariance term may be small
  - first two terms could be estimated directly, but would require cumbersome computations for each y
- **Stratified delete-a-group jackknife** for A only
  - use earlier $A \cup B$ jackknife weights to compute replicate control totals
  - calibrate each set of A-only replicate weights to the replicate controls

Summary

- Principled weighting methodology for combining probability and nonprobability data, accounting for sampling design and differential propensities
- Calibration strategy for producing microdata set using only the probability data source
- Replication-based variance estimation at all levels
- Questions or comments welcomed:

breidt-jay@norc.org

- Thank you!