National Cancer Institute

Strategies to Increase Access to Small Area Data and Resource – The NCI Experience

Benmei Liu, Ph.D.

Statistical Research and Applications Branch,
Surveillance Research Program,

Division of Cancer Control and Population Sciences,

NCI/NIH

NCVHS September Full Committee Meeting

September 13, 2018

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

I. Current activities underway to improve access to county and sub-county level

Introduction

- Cancer-related measures pertaining to risk factors, screening, policies, and knowledge are of great interest to cancer control planners, policy makers, and researchers at the state and county levels.
- Accurate local statistics have often been difficult to obtain.
- The standard direct estimates (design-based) from national survey data cannot provide reliable estimates due to the small sample size or are not available due to zero samples.
- Model-based small area estimation (SAE) methods that combine information from multiple related sources have been developed to increase the precision

Overview of the Model-based SAE Techniques

- ***Borrowing strength*** from relevant sources (Census/ Administrative information, related surveys)
- Borrowed strength comes from covariates, and from other counties with similar characteristics
- Methods of combining Information
  - Choose good small area model
  - Use good statistical methodology
- Mixed models (fixed effects + random effects) at area level or unit level have been popularly used in the small area estimation literature (Rao and Molina 2015, Jiang and Lahiri 2006).
- Among the many models developed in the SAE literature, the most prominent approach is the ***Fay-Herriot*** area- level model, originally developed to estimate per-capita income for U.S. areas with populations of less than 1,000.

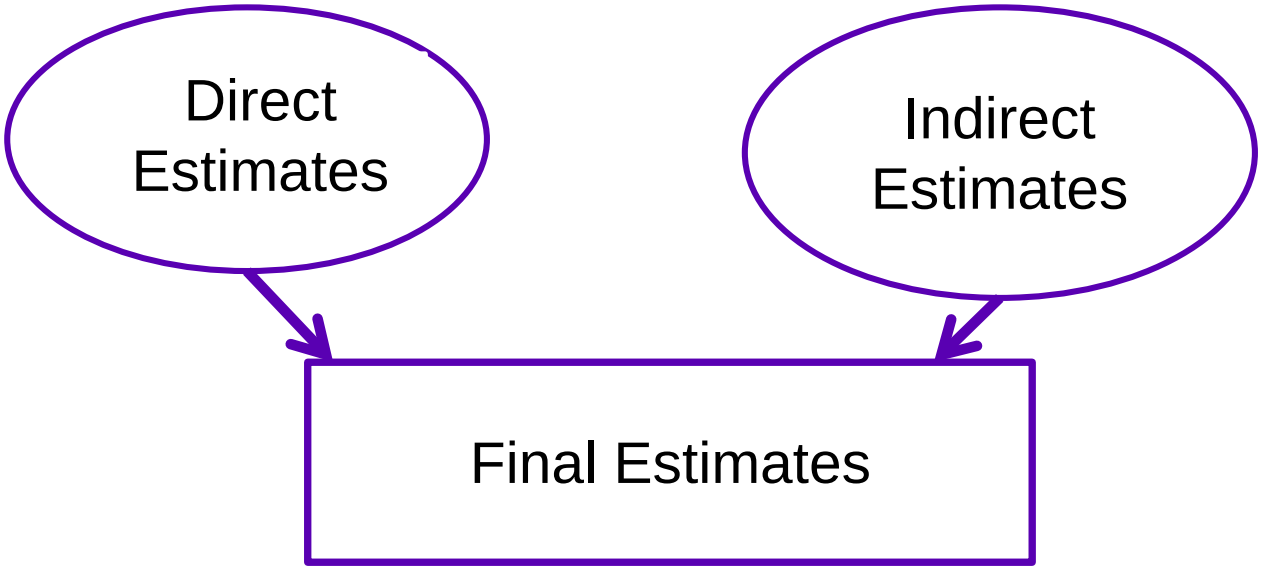
Fundamental Area Level SAE Model: Fay-Herriot Model (Fay & Herriot 1979)

- Sampling model: $Y_i|\theta_i \sim N(\theta_i,\ D_i)$;
  - Y_i is the direct survey estimate of the small area mean θ_i
  - D_i is the sampling variance and is typically assumed known
- Linking model: $\theta_i = X_i'\beta + v_i$; where $v_i \sim N(0, A)$;
  - X_i denotes a set of area-specific predictors
  - β and A are unknown model parameters

✓ Several transformations on the direct estimates Y_i are proposed to stabilize sampling variance D_i.

Statistical Inferences Using Mixed Models

- The final estimates are combinations of the direct estimates with the synthetic estimates.



- Fully Bayesian approach or empirical best prediction approach (analytic formulas) can be used for the estimation.

Current Small Area Estimation Projects at NCI

- Small area estimates using the NCI-sponsored Tobacco Use Supplement to the Current Population Survey (TUS-CPS)
  - County level estimates for two data cycles are produced for five tobacco related outcomes (https://sae.cancer.gov/tus-cps/)
  - Collaboration between NCI and Census
- Small area estimates using the NCI-sponsored Health Information national Trends Survey (HINTS)
  - State level estimates are produced for 15 cancer-related knowledge variables (https://sae.cancer.gov/hints/)
- Combining BRFSS/NHIS for Cancer Risk Factors and Screening Behaviors at the State and County Level
  - https://sae.cancer.gov/nhis-brfss/
- Spatio-temporal Models for Cancer Burden Mapping
  - To estimate age-standardized incidence rates by US county from a number of cancers and map the estimates to identify patterns and outliers.
  - Collaboration between NCI and American Cancer Society

Combining BRFSS/NHIS for Cancer Risk Factors and Screening Behaviors at the State and County Level

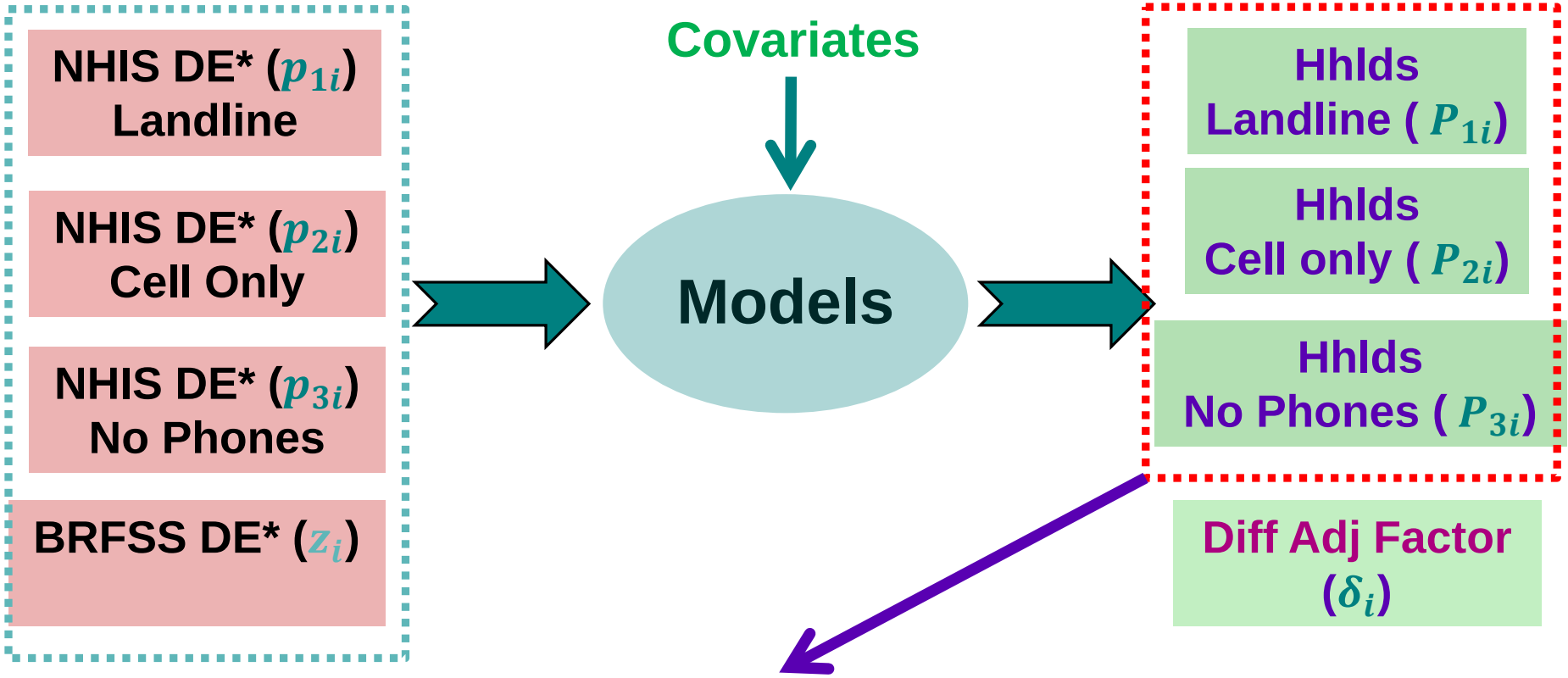
Collaborators for the data periods up to 2010:

- Eric Feuer at NCI
- Dawei Xie, Qiang Pan, University of Pennsylvania
- Van Parson, Nathaniel Schenker, National Center for Health Statistics
- Trivellore Raghunathan, University of Michigan School of Public Health
- Michelle Town, National Center for Chronic Disease Prevention and Health Promotion
- Information Management Services

Surveys Used

- Behavioral Risk Factor Surveillance System (BRFSS) – the largest U.S. survey tracking health conditions and risk behaviors at the state and sub-state level since 1984
  - + Large; almost all counties in sample
  - - Telephone survey
    - →Non-coverage of non-telephone households (only samples cell phone only households in recent years)
    - →Low response rates
- National Health Interview Survey (NHIS) – the principal source of information on the health of the civilian noninstitutionalized population of U.S. since 1957
  - + Face-to-face survey
    - → Includes non-telephone households (and a question identifying phone status of households)
    - → High response rates
  - - Smaller; only about 25% of counties in sample.

Statistical Models and Inferences for the Two Newer data Periods (2004-2010)



- **Final Estimates** are weighted summations of the three components classified by phone status: $P_i = M_{1i}P_{1i} + M_{2i}P_{2i} + M_{3i}P_{3i}$,

where M_{hi} &$P_{hi}, h = 1,2,3$ are the estimated telephone rates and proportions for the binary outcome of interest in each small area i for a specific time period, $\sum_{h=1}^{3} M_h$=1.

*DE: Direct Estimate

Project Achievements & Current Activities

- Bayesian methods are developed to combine information from the two surveys; also incorporated telephone coverage rates estimated from the census or NHIS
- Developed estimates for four time periods: 1997-1999, 2000-2003, 2004-2007, 2008-2010
  - Smoking, mammography, pap smear, Colorectal
  - Counties, health service areas, and states
- Estimation for years 2011 and forward is in progress
  - New collaboration with NCHS and CDC
  - Refining the outcomes, covariates and the methodology

Project Dissemination

- Publications:
  - Methods paper: Raghu et al (2007 JASA)
  - Application paper: Davis et al (2010 Public health Reports)
  - New manuscript: Liu et al (2018) is under journal review
- Websites:

  sae.cancer.gov

  http://statecancerprofiles.cancer.gov/

  Http://seer.cancer.gov/seerstat/variables/countyattribs/
- Communicating with end users
  - Conducted two focus groups with cancer control planners and public health professionals
  - Email communications

Usefulness of the Small Area Estimates

- Becomes an important data resource for cancer research
  - Charles Harding et al. (2015 JAMA) uses the SAE results for mammography and examined the relationships between mammography, breast cancer incidence and mortality at the county level. The findings have direct relevance to overdiagnosis, suggested that it is widespread.
- Motivates demands/new research on small area estimates for additional outcomes (e.g., prostate cancer screening prevalence)

II. Additional strategies are being developed/proposed in the near and long term

Improving Health Research on Small Populations

Challenges in conducting health research that is representative and informative:

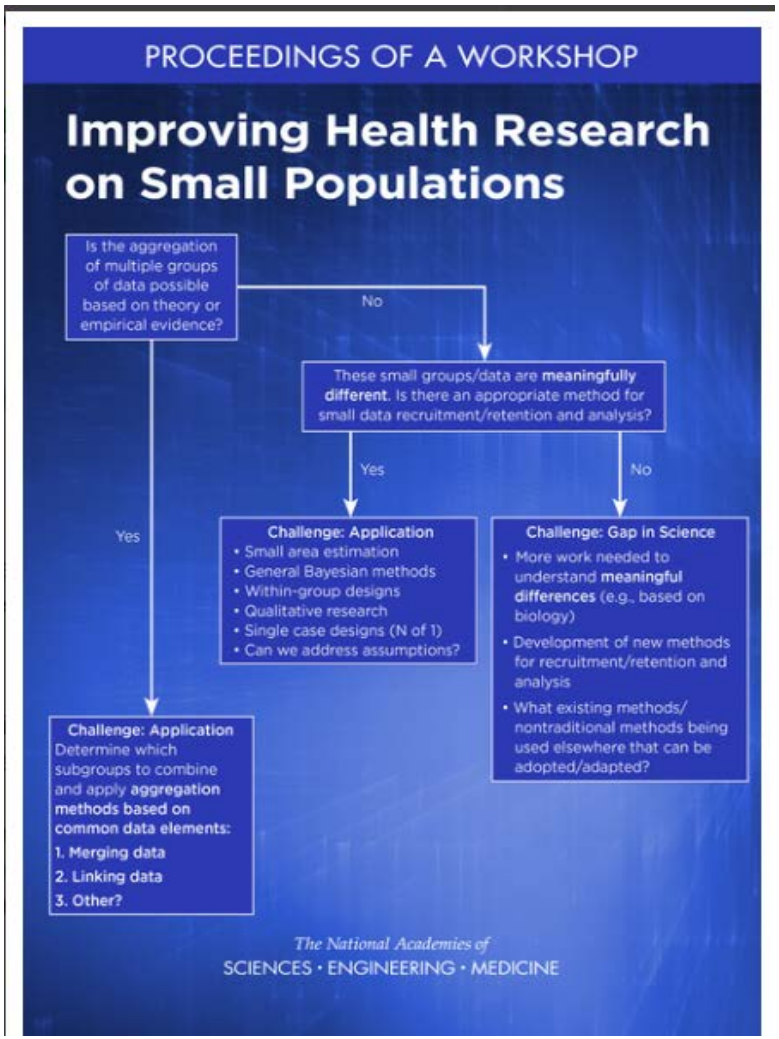
- Dispersion and accessibility issues can increase logistical costs
- Difficult to obtain adequate sample size - likely to be expensive
- “Meaningfully different”

Workshop in January 2018 by NASEM to discuss

- Alternative study designs
- Innovative methodologies for data collection
- Innovative statistical techniques for analysis

✓ **Dr. Shobha Srinivasan at NCI is the lead for promoting this agenda for small population work**

Report from the Workshop



http://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_181256

III. Application of New Technologies such as Synthetic Data to Provide Health Information Without Threatening Privacy

Applications of Synthetic Data

- Research on releasing synthetic census tracts in cancer registry data while maintaining confidentiality (lead by Dr. Mandi Yu)
  - Yu M, Reiter JP, Zhu L, Liu B, Cronin KA, Feuer EJR (2017). Protecting confidentiality in cancer registry data with geographic identifiers. *Am J. Epidemiology*, 186(1): 83-91.
- Using representative synthetic data to evaluate record linkage software
  - Liu B, Yu M, Feuer EJ. Evaluating record linkage software using synthetic data. Presented at the 2018 *NAACCR annual conference*, Pittsburg, PA.

Applications of Other Approaches

- Releasing specialized cancer registry database upon request with pre-calculated census tract based socioeconomic quintiles and two census tract-based rurality indicators
  - Yu M, Tatalovich Z., Gibson JT, Cronin KA (2014). Using a composite index of socioeconomic status to investigate health disparities while protecting the confidentiality of cancer registry data. *Cancer Causes Control*, 25(1): 81-92.
- Releasing census tract level neighborhood socioeconomic status index and walking related environmental factors (created using geospatial techniques) to the public use data to facilitate neighborhood level analysis
  - GeoFLASHE, a geospatial extension of the Family Life, Activity, Sun, Health and Eating Study (FLASHE) Https://cancercontrol.cancer.gov/brp/hbrb/flashe.html
- Similar technology maybe considered by other surveys such as the Health Information National Trends Survey at the zip code level.

Summary and Discussion

- The model-based SAE techniques represent an effective means of generating estimates where there is small (or zero) state or county sample.
- The SAE results, which are released and disseminated at several NCI's websites provide a useful resource for the broad cancer surveillance society to fulfill multiple needs.
- The small population meetings and publications are promoted as part of the rural cancer control research agenda.
- New techniques such as synthetic data and composite index could provide efficient ways to provide health information without threatening privacy.

Thank you!

Acknowledgments:

Dr. Laura Dwyer, BRP/DCCPS/NCI

Dr. April Oh, BRP/DCCPS/NCI

Dr. Shobha Srinivasan, OD/DCCPS/NCI

Dr. Mandi Yu, SRP/DCCPS/NCI

Contact info:

Benmei Liu

liub2@mail.nih.gov