Measuring Racial/Ethnic Disparities via Indirect Estimation: An Overview of Methods

Marc N. Elliott, Ph. D.
Senior Statistician
RAND
elliott@rand.org

March 25, 2010
The Brookings Institution Conference:
Charting a Course for Health Care Quality Improvement
Coauthors

This is joint, ongoing work with

– Allen Fremont (RAND)
– Nicole Lurie (HHS)
– Peter A. Morrison (RAND)
– Dan McCaffrey (RAND)
– Philip Pantoja (RAND)
– David Klein (CHB)
Self-Reported Race/Ethnicity:
Important, but often Unavailable

- To address health disparities, we need to classify populations by race/ethnicity (R/E)
- Self-report is considered the gold standard
- But self-reports are often unavailable and infeasible to obtain quickly
  - Most plans collect enrollee’s name & address, but not R/E
  - At best, self-report available only for a fraction of health plan enrollees
  - CMS Administrative race/ethnicity is known to undercount Hispanics, Asians
  - Other public records that might be valuable for research lack R/E information (but include name and address)
Health Plans, CMS, Researchers want to Measure Health by R/E

Aims include:

- Improvements in
  - Equity
  - Overall Quality
- Targeted interventions
Benefits of Indirect Estimates of R/E

• Timely, efficient, and inexpensive way to:
  – Understand R/E composition & distribution of patients served
  – Estimate R/E disparities in care & outcomes

• Fill gaps as self-reported R/E data accrues
  – Examine potential selection bias in initial self-reported data

• Link to GIS mapping and decision tools
Name and Address Methods

- **Surname lists**
  - Dichotomous surname lists with high sensitivity & specificity already exist for Hispanics, Asians
  - Census Bureau’s new 6-category (OMB) surname list gives R/E distribution of all 151,671 names appearing 100 times or more in 2000 Census (Word et al. 2007)

- **Geo-coding census information by residence address**
  - R/E distributions are published at the block group level
Hybrid Approaches

- Integrate complementary strengths of surname lists & geocoding
  - Surname lists are best at distinguishing *Asians* and *Hispanics* from others
  - Geocoding is best at distinguishing *Blacks* from others

- Hybrid approaches differ according to:
  - Information sources (e.g., new vs. dichotomous surname lists)
  - How they combine information
    - Sequential triage (Fiscella and Fremont 2006)
We Currently Recommend the Bayesian Improved Surname Geocoding (BISG) for General Purposes

- Combines block group racial/ethnic info from Census SF1 file
- Incorporates Census’ latest surname list probabilities (Word et al.), plus fixes for unlisted names, suppressed counts
- Produces a Vector of 6 probabilities of being Hispanic, NH White, Black, API, AI/AN, Multiracial
- Described in Elliott, Morrison et al. (2009, HSORM)
BISG Is Efficient for Hispanic, White, Black, API (4 Largest Groups)

- Strongest for Hispanic, then Asian, then White, then Black
  - C-statistic/area under the curve 92-98%
  - For many purposes 175 indirect cases=100 self-report
- Fairly similar performance by gender (and not always better for men)
- Not very useful for AI/AN, multiracial categories yet
- Has been extended to predict Spanish language preference (97% concordance)
- Could be extended to Asian and Hispanic subgroups
A Further Modification of BISG
updates CMS’
Administrative Race/Ethnicity Field

• Turns CMS Administrative classification into a vector of probabilities using previous self-reported data
• Independently calculates a vector of BISG probabilities from name, address
• Uses similar Bayesian approach to integrate these 2 estimates, 3 information sources.
• Substantially improves the original CMS variable
• Will be used for CMS reporting of CAHPS & HEDIS
Can Use BISG Probabilities for Interventions, Sampling

- **Interventions**
  - Flexible: Can choose any desired probability threshold
    - e.g., a letter to anyone more than 30% likely to be Asian
  - Most likely cases for a fixed budget:
    - Select 500 members most likely to be Hispanic

- **Sampling for Research Studies**
  - Can use probabilities to oversample certain racial/ethnic groups or improve power to compare R/E groups
  - Illustration: Used names/addresses on marriage license applications for a study of racial/ethnic differences in effects of marriage on health
We Have Developed Efficient Analytic Methods for Using Indirect Estimates

- Our primary focus has been on maximizing accuracy of disparity estimates
  - Best estimates of race/ethnic distribution, best estimates of individual race/ethnicity, and best estimates of disparities are three different things
  - Only meet at 100% accuracy; otherwise can be tradeoffs
- Main Recommendation: Don’t dichotomize/classify race/ethnicity for analysis
  - Categorizing individual race/ethnicity before estimating population disparities needlessly loses information
  - Can also bias estimates
- Use the probabilities directly, interpret as if were categorical
- Can do anything that you can do with categorized data, but more efficiently
  - See McCaffrey & Elliott (2008, HSR)
  - We’ve developed sample SAS codes for this (pantoja@rand.org)
- Can combine indirect and direct data as self-report accumulates
Conclusions

• Bayesian combination of surname and address info is a powerful new means to infer race/ethnicity from administrative records
  – Can be integrated with incomplete administrative info and partial self-report
  – Can predict language preference
• Can be used for targeting interventions, sampling
• Analytic approaches use probabilities directly
• Links to geography (and hence GIS) opens the door to further insight about contributing factors at local level
Support

• Robert Wood Johnson Foundation
  – Brookings-RAND Racial/Ethnic Healthcare Equity Initiative

• CMS
  – RAND HHSM-500-2005-000281
