ENHANCING DATA SHARING VIA "SAFE DESIGNS"

Generating Knowledge To Inform Scientific Practice

Kristine Witkowski

Inter-university Consortium for Political & Social Research, University of Michigan

DATA SHARING CONTEXT

- U.S. policy requires submission of data sharing plan, when applying for research funding
- Current effort to revamp process for protecting human subjects (ANPRM 7/22/2011)
- Multifaceted approach when formulating data for safe and optimal use (Lane 2007)
- Need to think about data sharing early and often, using specialized knowledge

GUIDING PRINCIPLE

- Producers must be able to effectively draw upon disclosure research to accurately determine the work required to optimally meet data sharing goals

AIM

- Enhance the value and safe use of social science data – particularly for contextualized microdata
- Simulate scientific practice to generate knowledge for broad and responsive use

RESEARCH PROJECT

- 5-year project supported by National Institute for Child Health & Development
- Dan Brown University of Michigan
  Michael Elliott
  Trivellore Raghunathan
  Kristine Witkowski
  Kevin Leicht University of Iowa

ADVISORY BOARD

- John Abowd, Cornell University
- Marc Armstrong, University of Iowa
- Jerry Reiter, Duke University
- Natalie Shlomo, University of Southhampton
- Christopher Skinner, London School of Economics & Political Sci.
- Laura Zayatz, U.S. Census Bureau

DISCLOSURE SIMULATIONS

- Simulate disclosure work for representative series of artificial microdata files
- Estimate disclosure outcomes, as measured for a comprehensive set of risk, utility, and cost elements
- As determined by alternative specifications of sampling and database design parameters
- Controlling for iterative sets of survey-sites (or a specific set targeted for collection)

DISCLOSURE SIMULATIONS

- Restricted microdata from the American Community Survey provides geographically-specific information used throughout project
- Artificial files offer methodological flexibility as well as data confidentiality
- Project conducts experiments to assess the accuracy of estimates derived from artificial data

MODELS FOR ARTIFICIAL DATA & POPULATION REIDENTIFICATION PROBABILITIES

- Estimate composition of likely-participants as well as general study population
- Multiple imputation
- Joint probability distributions for 1-km^2 pixels
  - Identifying personal attributes and non-identifying health outcomes
  - LandScan, decennial census, ACS microdata, BRFSS
  - Areal weighting methods to estimate pixel data from more aggregate data (i.e., blockgroups)
  - Controlling for non-response (weighted vs. unweighted)

METADATA

$$\mu^m_a\,;\,\sigma^m_a\,;\,\delta^m \;=\; f\,[\,s, r, d\,]$$

For any given disclosure outcome (m) resulting from sample (s) , release (r), and SDL (d) design elements as estimated from replicating artificial files (a, f)

Where:

μ^m_a = Estimated outcome (mean)

σ^m_a = Variance of estimated outcome (reliability, precision)

δ^m = Difference from observed outcome (validity, accuracy)

$$o^{m}_{ra,f} = m(o^{m}_{--,-}) + m(o^{m}_{ra,-}) + e(o^{m}_{ra,f})$$

Where:

f = File as compiled from specific sample iteration

ra = Experiment using either real (r) or artificial (a) data

m = Different measures of disclosure outcomes

$o^{m}_{ra,f}$ = Disclosure outcome for file

$m(o^{m}_{--,-})$ = "Grand" mean outcome across all files

$m(o^{m}_{ra,-})$ = Mean outcome for real or artificial files

$e(o^{m}_{ra,f})$ = Variation among real or artificial files

Accuracy of estimated outcome

F_o^m = MST (Between) / MSE (Within)

$\delta_\mu^m = [m(o^m_{a,-}) - m(o^m_{r,-})] / m(o^m_{r,-})$

$\delta_\sigma^m = [s(o^m_{a,-}) - s(o^m_{r,-})] / s(o^m_{r,-})$

$\phi^m = s(o^m_{r,-}) / s(o^m_{a,-})$

$\theta^m = m(o^m_{r,-}) - [\phi^m * m(o^m_{a,-})]$

Estimated outcome (adjusted)

$$\mu^m_a = E(\theta^m) + [E(\phi^m) * m(o^m_{a,-})]$$

Variance of estimated outcome (adjusted)

$$\sigma^m_a = E(\phi^m) * s(o^m_{a,-})$$

METADATA

$$\mu^m_a ; \sigma^m_a ; \delta^m \quad = \quad f\,[\,s, r, d\,]$$

For any given disclosure outcome (m) resulting from sample (s) , release (r), and SDL (d) design elements as estimated from replicating artificial files (a, f)

Where:

μ^m_a = Estimated outcome (mean)

σ^m_a = Variance of estimated outcome (reliability, precision)

δ^m = Difference from observed outcome (validity, accuracy)

SAMPLE ELEMENTS (s)

- Study population of adults (age 18 +)
- Limited study region: Indiana, Illinois, Michigan, Ohio, Wisconsin
- Household survey based on two-stage sample of tracts and housing units clustered within
- Total sample size
- Detailed sampling design – locations, target populations, and sampling rates

RELEASE ELEMENTS (r)

- Person-Level
  - Identifying characteristics of respondent (e.g., age, sex, race/ethnicity, obesity-status, household composition, spousal attributes)
  - Non-identifying health outcomes: Self-reported health, chronic condition (e.g., diabetic)
  - Sets of 6 or 10 attributes, held constant

RELEASE ELEMENTS (r)

- Geography-Level
  - Direct identifiers of region, state, & population density (e.g., MSA-status)
  - Indirect identifiers or contextual variables
    - Administrative and georeferenced spatial-units: Counties, tracts, blockgroups, & 1-km^2 pixels
    - Public-use data: Census, EPA, NASA, others
    - Sets of variables of broad interest (wishlists)
    - Samples representative of all possible sets

RELEASE ELEMENTS (r)

- Geography-Level
  - Indirect identifiers or contextual variables
    - Domain or measurement: Population and housing characteristics, air quality, tree coverage, proximity to incinerators, miles of road
    - Type or areal size of underlying geography: Pixels, blockgroups, tracts, & counties
    - Number of variables to be released
    - Entropy

SDL ELEMENTS (d)

- Linkage Experiments: Geographic-Level
  - Strangers and acquaintance intruders
  - Link to public sources of contextual variables
    - Complete and accurate data
  - Matches: Geographies (in population) with same attributes as surveyed locations
  - Blocks: Region, state, population density
  - Personal attributes, coupled with geographic attributes, used to refine estimates that particular areas have been drawn into study

SDL ELEMENTS (d)

- SDL Techniques: Geographic-Level
  - Assume personal identifying variables are not masked
  - Applied after collection: Global recoding and synthetic values of contextual variables
    - Deterministic linkage, probabilistic linkage, k-nearest neighbor, Mahalanobis distance, others
  - Applied before collection: The “Safe Design”

SAFE DESIGN

- Formulate innovative SDL technique for addressing reidentifying personal attributes, holding constant geographic attributes
- Study that supplements their sample and responsively collects data to minimize risk of being a sample unique (i.e., k-anonymity)
- Circumvents constraints from established practice of addressing disclosure after data are collected

SAFE DESIGN

- **Baseline** sample: Sampling design formulated to meet analytical goals (U_b, C_b)
- **Preemptive** disclosure review: Disclosure risk of baseline sample (R_b)
- **Supplemental** sample: Sampling design formulated to meet confidentiality goals ($R_s \sim 0$, $U_s > U_b$, $C_s > C_b$)

Where: R = Risk, U = Utility, C = Cost

DISCLOSURE OUTCOMES (m)

- Risk
  - Identity disclosure: Population reidentification probabilities and k-anonymity
    - Persons in study population sharing similar geographic and personal attributes
    - Respondents sharing similar geographic and personal attributes within data release
  - Continuous cell sizes; at-risk status with thresholds defined by content sensitivity
  - Per record – per target subpopulation – per design

DISCLOSURE OUTCOMES (m)

- Utility
  - Information loss: Characterizing release as a whole, including both continuous and categorical measures, scale-invariant
    - 12 measures provided by Domingo-Ferrer, Torra, and Mateo-Sanz
  - Suppression bias: Geographies and subpopulations most at-risk
  - Statistical inference: Relationships between health outcomes and spatial contexts

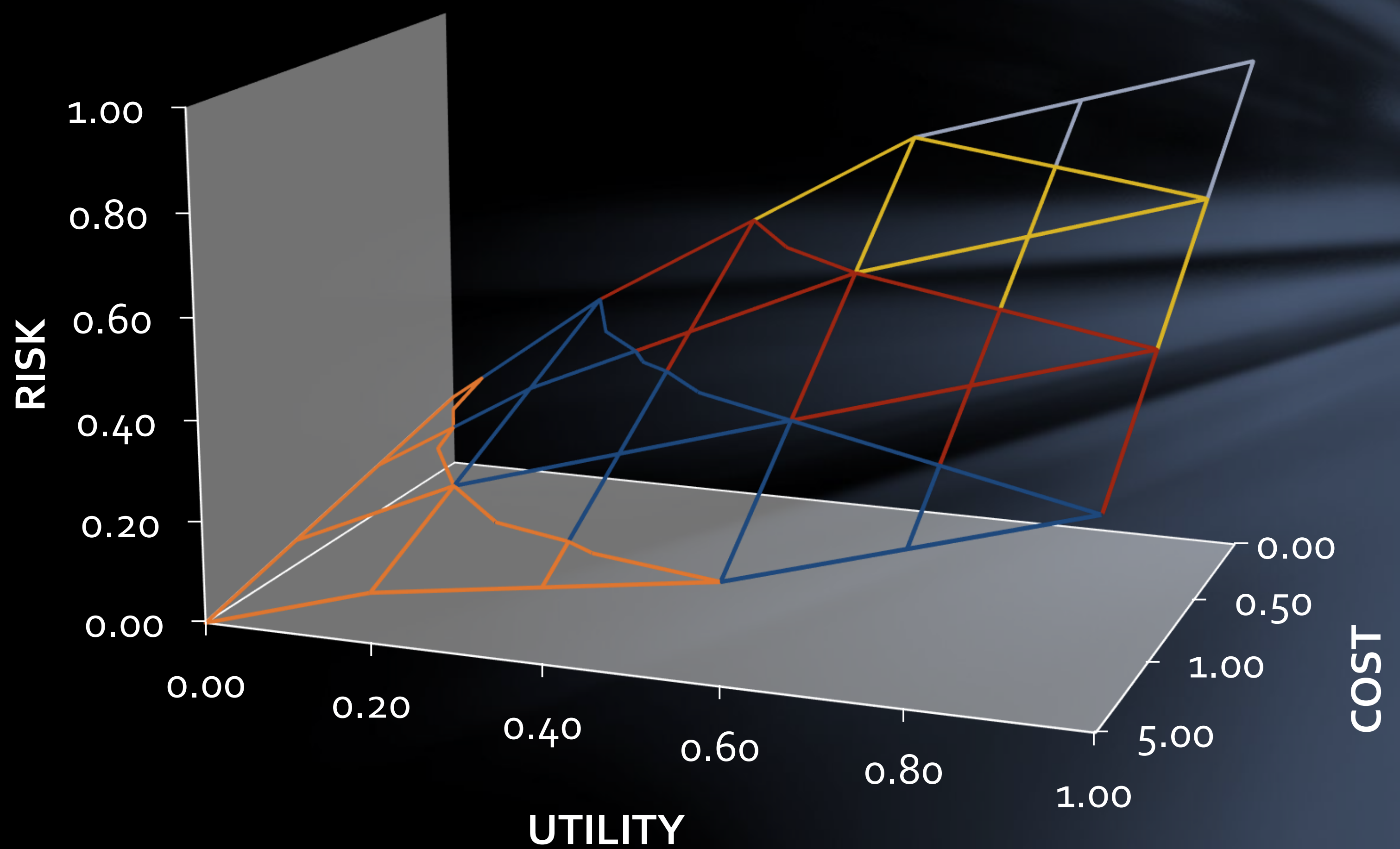
DISCLOSURE OUTCOMES (m)

- Cost
  - Average dollar values of survey expense
  - Function of number of draws required to meet targeted sample sizes for broadly defined and detailed subpopulations
  - Directly informed by scientific practice

ADDITIONAL CONSIDERATIONS

- Added value and cost of spatially-dispersed samples that maximize variance in geographic attributes (s)
- Trading-off data on personal attributes for geographic detail (r)
- Protection offered by measurement error and concentration of hard-to-count populations (d)
- The role of administrative data sources (d)

RISK-UTILITY-COST MAP



IMPLICATIONS

- Flexible framework for generating empirical data that can broadly inform decision-making
- Supports sharing and consumption of complex and highly specialized knowledge
- Supports policies regarding data sharing and protection of human subjects
- Audiences: Established and new studies of federal statistical agencies and academic institutions; DRBs, IRBs, archives; funders

THANK YOU. QUESTIONS?