Administrative Record Quality and Integrated Data Systems

Robert Boruch

University of Pennsylvania

November 28-29, 2011

ISP Conference

Aims

- Identify, conceptualize, and illustrate issues on data quality that are important in…
- Statistical analysis in IDS as opposed to making decisions about individuals in IDS
- And to discern gaps in our understanding

Theme

- Really, universally, relations stop nowhere, and the exquisite problem of the artist is eternally but to draw a geometry of his own, the circle within which they shall happily appear to do so (Henry James, 1907).
- The analyst in an IDS context might be said to have something like the same problem except that...

Assumptions

- IDS Function(s)
- Administrative Record vs. Research Record Function
- IDS Administrative and/or Research function
- Capacity to integrate, link, merge, etc. will continue to expand on account of tech.
- Emphasis on evidence based decisions/policy will continue to expand.

Resources Used Here

- Peer reviewed research reports, including government reports, on quality in…
- Health
- Crime and justice
- Education
- Economics
- National, sub-national, cross nation
- Not: intel fusion, private businesses, CS, SS

Resources on Topic

- Sparse: Very few comprehensive across sector
- Fragmented often: Piecewise
- Unconnected: Disciplinary provincialism
- Nonetheless: Missingness, measurement error, etc. in occasional papers, EHandbook, Quality and Record Linkage, Survey methods.
- IDS Dimensions: Temporal (RCTS), jurisdictional, hierarchical, domain, quality indicators/data

Provincialism

"Existing research results show that researchers (on data quality) are primarily operating in two major disciplines-Management Information Systems and Computer Sciences." ACM Journal of Data and Information Quality (2009).

Background and Definitions

- IDS: Putting records from A together with records from B, C, etc.
- Subject labels: Perp, client, customer
- Record content: Any medium (e.g. video)
- Integrated
- Variable and variable values
- Quality: Statistical error rubrics defined below.

Topical Checklist/Taxonomy

- Checklist as soft technology/practice guideline
- Checklist as a taxonomy
- Taxonomy as science
- Conceptual frame work, blah blah

Checklist/Taxonomy Ingredients

- Target population coverage
- Variable definitions, uniformity (harmony)
- Interpretation of definitions
- Jurisdictional variation in definitions
- Distortions: Systematic, local etc.
- Distortions: Random, local etc
- Missing data: Systematic & random
- Local theories of distortion, missingness
- Incentives
- Prospective audit studies
- Tolerable error

Philosophy of Science

"The fact of it is that knowledge can only advance over a battlefield strewn with eliminated errors." Rescher (2007)

"Hyperbolic, but engaging. Now we get to the categories of error." (Boruch 2011)

Target Population Coverage

- Examples: Uniform Crime Reports (UCR) on rape at .3/1000 vs National Crime Victimization Survey (NCVS) at .5/1000, and prop crimes at 34 vs 154/1000.
- More generally, surveillance systems and spontaneous reporting, e.g. FAA near-crashes, FDA post-marketing surveillance, vaccine coverage in African countries, etc.
- Reco: Capture-recapture statistical methods and IDS
- Reco: Levels of linkage: Individual, location based, etc.

Theme: Labels, Relabels, Etc.

- One of the unpardonable sins in the eyes of most people, is for a man to go about unlabeled. The world regards such a person as the police do an unmuzzled dog, not under proper control (T.H. Huxley, 1893).
- Example: Race/ethnicity nowadays versus old days

Variables and their Definitions

- Identity: DNA, SSN, names plus, etc., etc.”David Wilsons”
- Activity, e.g. misdemeanor/felony assault with a knife
- Status e.g. homelessness, death, poverty level, income
- Recos: make feasible adjustments, focus on percentages and trends within and across systems rather than absolute numbers, drop really problematic variables.

Theme: Left Out Variables

- Measurement does not necessarily mean progress. Failing the possibility of measuring that which you desire the lust for measurement, for example, merely results in your measuring something else – and perhaps forgetting the difference – or in your ignoring some things because they cannot be measured (George Udny Yule, 1871).
- Witness the use of federally subsidized school lunch programs as an indicator of poverty in conventional linear models based analyses as opposed to mother's IQ or education.
- Charles F. Manski (1995) Identification Problems in the Social Sciences. (2007) Identification for Prediction and Decision.

Interpretation of Variables

- The subject who reports, e.g. Aboriginal students in Australia.
- The categories for subject reports, e.g. U.S. Census 2010 versus earlier years.
- Categories for recorders' reports, e.g. death certificate and entries on gender and age.

Theme: Language and Records

- With this booklet, all I had to do was run my finger down the left column until I found the English phrase I wanted, and then rattle off the nonsense syllables printed opposite in the right hand column.
- “How many grenade launchers have you?,” for instance, was “Vee feel grenada vairfair habbenzee.” Kurt Vonnegut (1999).

Jurisdictional Variations and Interpretation

- State-wise variation, e.g. Federal NDEA definitions of disabilities and state discretion. Autism Spectrum.
- "…in surveying 50 states, Murphy found 'something like 40 different terms to describe the rape of a child.'" and FBI on rape index crime definition (Brisbane, 2011, NYTimes "Confusing Sex and Rape"
- Municipality/local variation, e.g. misdemeanor domestic violence
- Nation-wise variation, e.g., EU and crime type
- Nation level/Institution-wise, e.g. graduation

Theme: Random Error Maybe

- It is the supreme law of unreason whenever a large sample of chaotic events are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along (L.H.C. Tippett, 1943).

Distortional Influences That Are Random, Maybe

- Statisticians like random error, if there must be any error at all, e.g. 300 year history
- Good evidence of random error in certain measures, e.g. ed achievement, psych, physics
- Evidence that error in administrative records is random seems sparse as yet.
- In criminological research: classic random error model is distrusted for self reports/surveys and for administrative records

Systematic Distortional Influences

- Police: Cop A likes to arrest people. Cop B does not.
- MDs: Doc A likes to prescribe pills. Doc B does not.
- Service Providers: Staffer A likes to "help" families into Head Start. Staff B abides by rules.
- Recos: Not necessarily a quality of record issue. BUT, check reasonableness of entries, have someone check sample of data, have a second person enter data and compare, train or retrain, periodically audit.

Theme: Independence of Sources

- “It is ever so hard, when a concrete fact illustrates a hope, to weigh that fact properly. When the first six people we meet agree with us, it is not so easy to remember that they may have read the same newspaper at breakfast” (Walter Lippmann, 1922).
- Standards for sainthood. (Kruskal JASA).
- Reco: Get independent sources when possible, and ask are they really independent.

Missing Data

- For the statistician, the best way of handling missing data is not to have any (Lincoln Moses circa 1985). Missingness will always be with us (Boruch circa 1985)
- Example: Use of software in medical teamwork to reduce missingness, once a drug is ordered (corollary orders)
- Example: Police records and male victims of assault, 20% missing data.
- Example: Banks and 60% complete in credit risk DB
- Recos : Try to fill in the gaps, exclude missing or estimate, terminate or alter the study.

Theme: Systematic Distortion

- “The more any quantitative social indicators are used for social decision-making, the more subject it will be to corruptive pressure and the more apt it will be to distort and corrupt the processes it is intended to monitor” (Don Campbell, 1975).
- Anecdotes are in ample supply. Coherent and evidence based approaches to understanding this are not except for Ibn Khaldun, Knightly.

Local Theories of Distortion, Missingness, and Quality of Record in an IDS

- Old news, perhaps: A jurisdiction's use of optical scanning is better than manual transcription (Abowd and Vi Huber 2004 on California).
- But, not much transparent evidence on comrade Campbell's declaration.

Theme: Provenance

"The Government are very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root, and prepare wonderful diagrams. But what you must never forget is that every one of these figures comes…from the *chowty dar* (village watchman) who put down what he damn pleases." Stamp (1929)

Provenance and When: Incentives

- “A CIA review conducted in 2005… found that as the possibility of war increased, intense interest in… Iraq’s WMD capabilities lowered the threshold for reporting real information and increased the volume from less credible sources” (US Senate Select Committee on Intelligence, 2006).

Provenance, When, and Why: Incentives

- In medicine, reimbursement systems lead to distortion in coded diagnoses in medical records, e.g. 80% match between charts and reimbursement (McCarthy et al. 2000).
- “Suicide” as a social construction, and reports on them, e.g. Whitt (2006) on NYC declines in rate during 1985-1989 with a new medical examiner.

Disincentives

- "… it must be recognized that people do not like filling in forms. The farmer is interested in growing and tilling crops, and he regards the making of a statistical reform as a pestiferous waste of time" (L.H.C. Tippett, 1943).

Reco

- Tag source and time of entry routinely (software can do this stuff now, we could not do in olden days)
- Look for local incentives to distort, create missingness, etc.
- Look for local disincentives to comply, distort etc.
- Study the matter

Prospective and Retrospective Audit Studies

- Prospective: counterfeit clients, patients, etc.
- Standardized patients in medicine, e.g. Peabody (2004) report 57% correct primary diagnosis.
- Standardized applicants for Head Start Services, e.g. GAO (2010) report 8 of 15 cases fraudulent misrepresentation.
- Verbal autopsies and malaria

Processing Administrative Records

- Resources have not been sufficient to reconnoiter this topic.

Theme: Linkage/Harmonization

- “A homphorism is a ‘transformation of one set into another that preserves in the second set between members of the first set.’ If this makes you glad that you are not a mathematician, I suspect you are not alone” John Henshaw (2006).

Administrative Record Linkage

- The quality of linkage of administrative records in an IDS depends heavily on the quality of the records' data.
- Linkage quality may be indexed and studied in a variety of ways, e.g. false positives, false negatives, recall rate, in deforministic or probalistic frameworks.
- This linkage topic is beyond the scope of this paper. But see Herzog et al. (2007).

Standards and Standardization

- NCES “Common Data Standard Initiatives”
- NCES “Standards for Statewide Longitudinal Data Systems (SLDS)” November 2010 Brief 2.
- OMB (1980)
- Others.

How Much Quality For What Purpose?

- Medieval Jewish History
- Medieval Arabic History
- Example: Teacher value added estimates from IDS of teacher and student records (mom's education).

Dual Mission Agencies

- Two functions: Enforcement and Statistical Characterization
- E.g. Police Departments, FAA, FDA, etc. etc.
- Not well documented.