Administrative Record Quality and Integrated Data Systems

Robert F. Boruch

Center for Research in Education and Social Policy
Graduate School of Education
University of Pennsylvania

Contents

> It is a mistake to think, as some do, that inaccurate or unreliable figures should not be given careful treatment. They may not merit it, but they certainly need it. Tippet (1943)

Introduction

This paper considers the quality of records in the context of Integrated Data Systems (IDSs). Here, IDS refers mainly to administrative records collected and maintained by different government agencies or non-government organizations (NGOs) on the same identifiable individuals. At times, the term IDS may refer to administrative records that are combined with records from a probability sample survey or records that are combined in the context of a controlled randomized trial. In an IDS, the records from the different agencies on the same individuals are put together. That is, the records from different sources are integrated into one data system.

The focus in this report is on integrated record systems in human and social services broadly defined. The military, intelligence, and engineering sectors, at this writing, use the phrases "information fusion" and "fusion centers" for related but different ends. These efforts in these sectors engender some of the same problems that social services IDSs encounter. The fusion initiatives are not considered in what follows.

The records considered here are ones which were initially constructed to inform the agency's administrative decisions about the individuals on whom the records are created. The purposes of the individual's record in each of the different agencies that contribute to an IDS will then often differ. Adult court records, for instance, will differ in purpose and character from records maintained by a hospital on the same person's visits to an emergency room. Birth and early childhood records on the person have different functions and may have a different name for the person than ones given in adult records.

The integration function in an IDS is important. Integration of individual's records, at times, permits us to understand statistical relationships that might otherwise not be uncovered by relying on a single record source. For instance, one may learn how individuals migrate from one foster home or school to another, and how levels of instability in early childhood antecede important events in later life by combining social service records with education records or juvenile court records. This, in turn, can help in understanding how different social systems and agencies may influence children's trajectories and better estimate the effect of special programs on the

individuals' health and well being. From an integration of occupation records with injury or health records at the local or corporation level, one may discern how different occupations pose health risks and whether prevention efforts to reduce injuries in various occupations work.

Further, any controlled randomized trial that is designed to test the effectiveness of innovations in the social sector can be construed as an example of IDS. The randomized trials called Moving to Opportunity, which Estelle Richman (2011) alluded to in her remarks at the ISP conference, required the integration of research records (who is in the control and intervention conditions) with administrative records from different sources. The reductions in obesity that were found in this experiment of moving poor people to more affluent areas were because of the ability to link records of children and their families (especially mothers) from different sources with research records in the randomized trial (Ludwig et al 2011).

Similarly, integration at a level higher than the individual may help in understanding and reducing problems. Anonymised patient records from identified emergency rooms on the person's location and nature of injury, for instance, may be shared and integrated with local police records to deploy violence reduction initiatives in the pertinent jurisdictions and to support evaluations of such initiatives, e.g. Florence et al. (2011). Records on families, however family is defined, may be integrated from different sources. We focus here on integration at the individual person's record level, although a few examples at higher levels are given in what follows.

The capacity to link records, to integrate, so as to understand the correlation between one variable and another, is remarkable, given contemporary information technology. The incentives to do so include professional interests in transcending conventional bureaucratic and academic discipline boundaries so as to advance societal interests. The incentives include the press toward evidence-based policy and practice, results-based policy, and similar initiatives in government.

Advocates of IDSs must nonetheless attend to whether the product of the integrated data system has value. We do not consider product value here. But we do keep in mind Henry James' (1907) remarks:

> "Really, universally, relationships stop nowhere, and the exquisite problem of the artist is eternally but to draw, by a geometry of his own, the circle within which they shall happily appear to do so."

This paper's aim is to identify, conceptualize, and illustrate key issues on data quality that are important in *statistical* analyses of the records assembled in an IDS, and to review research on resolving or understanding issues that one is likely to encounter when using IDS for policy analysis. The aim is also to identify gaps in our understanding of certain issues or topics such as how administrative functions of the records may lead to distortion or error in the record's contents.

In regard to statistical analyses, the statistician or social scientist attempts, for instance, to estimate the frequency of problems at the aggregate level, trends over time, or the relationships at the aggregate level among variables such as prognostic factors, e.g. predicting adult malfeasance from juvenile indiscretions. Or, the statistical aim may be to estimate the effect of interventions or services on groups of individuals or entities as in a controlled randomized trial.

The statistical uses of records assembled in IDSs are then apart from the uses to which the records on identifiable individuals are put by the agencies that generate and maintain the records. For instance, records of a welfare services agency permit the agency to make decisions about particular identifiable individual's compliance with agency rules based on the person's records. A judge may direct a particular juvenile offender to a diversion program, for instance, in an administrative action directed toward the individual. The same judge may engage in a randomized trial that, by lottery allocation, is designed to learn whether one diversion program works better than others, on average.

The functions of a research record, or of an administrative record used for research purposes, differ then from the normal administrative function of a record, i.e. making decisions about particular individuals. The difference is reflected in law and regulations, with processes and considerations discussed in Petrila's paper in this volume. Quality of the records' contents is of most concern here. The phrases "record quality" and "data quality" will be used interchangeably.

This report covers basic definitions and examples of how administrative record quality issues are explored. The main product is a topical checklist of items that ought to be considered. Standards for quality control, deciding how much quality is warranted for what purpose and at what cost, and dual mission (administrative and statistical research) agency missions are handled briefly. The implications for future research are identified. This report reiterates some of the practical advice given by people who have toiled in related vineyards to try to enhance record quality, and extends the advice where possible.

References to related work and acknowledgments are given, of course. Historical antecedents are intriguing but recognized only briefly. The emphasis is on systems based in the United States. Work from other countries is cited at times because the idea of IDS transcends geo-political boundaries.

Background, Definitions, and Resources

The idea of an IDS and its actualization has some recent origins. This report was invited in 2010, for instance, as part of the MacArthur Foundation's investments in a consortium on the topic of IDSs.

The idea has early origins in the social services sector, notably in efforts of the 1980s at Chapin Hall (Chicago), and later in South Carolina, Florida, Philadelphia, and a few other places, to assist agencies that maintain different kinds of records to capitalize on integrated data to improve social services. In the criminological and youth arena, Maltz (1996) was prescient in trying to understand whether and how and why to integrate probation, welfare, and health and housing code violations with juvenile records. Researchers at non-profit research organizations in the United States, such as the Urban Institute (Hatry, 2010) and the American Institutes for Research have also advanced the use of administrative records in evaluating social programs, and we rely on some of their work in what follows. The pedigree for IDSs also lies in work done during the late 1960s and 1970s in the United States, Canada, and the Scandinavian countries to integrate and employ administrative records from various sources in research that informs policy (Herzog, Scheuren, & Winkler, 2007; Janson, 2000).

DEFINITIONS

Definitions are fundamental in an IDS, as in any of the sciences. Moreover, the differences in the way that words are used in any particular data system within an IDS and in the academic disciplines are important.

The "subject" of a record is defined here as an individual person, whose unique identity can, in principle, be determined. The most unambiguous form of identification, one may argue, involves DNA, a kind of identification that is not, of course, easily obtained or recorded in many administrative records (police and health being partial exceptions depending on jurisdiction, etc.). But as of this writing, DNA is most accurate, and in future may be a part of agency records that constitute

an IDS. Depending on the administrative record system, the subject might be a family consisting of one or more persons, an entity such as a corporation, or an entire Girl Scout troop. We do not consider these alternative subjects of a record, i.e. clusters of persons, simply because these engender unique issues which cannot be handled at this time.

In the vernacular of a particular administrative record agency, the subject might be called a client or customer, "perp" or perpetrator, student/graduate, employee, patient, and so on. In the vernacular of controlled clinical trials, there has been a movement away from the use of the noun "subject" toward the word "participant."

For T. H. Huxley, writing in 1893, "… one of the unpardonable sins … is for a man to go about unlabeled." Be that as it may, in IDSs there must be some ways of assuring that the individual who is called a perpetrator in one kind of record system is the same individual identified as a customer, client, or patient in other record systems. Being able to handle these vernacular differences is important for anyone with responsibility for an IDS and for anyone in the operation who has to learn how to integrate well.

IDS glossaries that contain the vernacular definitions and relationships among them when they refer to the same subject are a natural way to handle this. Though individual IDS entities may construct glossaries for internal use, we have uncovered no publicly accessible glossaries of definitions and vernacular, as of this date, for the IDS context. The absence of such glossaries can be construed as a gap in our understanding.

A subject's "record" is defined here as an assembly of information in numerical, narrative, visual, audio, or other forms on the subject. Assume tentatively that the record is constructed and maintained continuously by an agency's record facility to serve the agency's purpose.

"Integrated records" on the same individual, the target subject of the record, means that different records on that individual from different entities or agencies are put together. In other scientific vernaculars, words such as "linked" or "merged" are used to designate the same product.

For statisticians and many social scientists, the word "variable" refers to a particular kind of entry into an administrative record that is relevant to the administrative agency's mission. Gender, for example, is a variable. The synonym for the word variable in some agencies might be "category," "characteristic," "item," and so on.

The variable as just defined may take on different "values." In different record systems, the value may be designated as an "entry" or "coded as," for instance. The possible values for indicating gender may be male (1) or female (2) or something else (3-6). In the criminological arena, prior convictions may be coded as 1 or 0 or more elaborately in any good police rap sheet. Race and ethnicity may now be understood in any combination of mixtures, at least in the United States in a federal context (Prewitt, 2005). A second generation child born of parents who were born in the United States may, in the U.S. Census, self-identify as both black and white, as Hispanic and non-Hispanic, and as Asian on account of their actual racial and ethnic origins. This opportunity to self-identify in more nuanced ways brings a mischievous glee to my graduate students, relatives, and others who do, indeed, have such mixed origins, but it presents serious challenges to the constructor of records and to the social scientists who want to get beyond conventional characterizations. This coding matter is not trivial for statisticians who want to understand how and by whom records are coded and what the coding means in the IDS.

"Error" here depends on the ways in which the survey statistician would define the word, e.g. Groves et al. (2009). At the level of population coverage in an administrative surveillance agency, for instance, error will mean a difference between what the intended target population actually is and what the administrative records at hand, on the same population of individuals, actually cover. At the level of an individual's report to an administrative agency for a record, error means the difference between the subject's report of the value on a variable, such as the subject's self-identification, and some higher standard such as a DNA identification or birth certificate. The error may be adjudged relative to a catenation of entries into different records, of course. The value of "date of birth" taken from a birth certificate, the value of birth date entered in a record on account of subject's response to a question about the date, a baptismal certificate, and parent's report may differ. Reconciliation rules then must be set up to take action on the error. We will exploit Groves et al. (2009) classification of errors in a statistical context in what follows.

RESOURCES

The peer reviewed research literature on integrating administrative records data and on the quality of administrative records that are integrated is ample in a few respects. It is sparse in many respects. Further, the dependable work on the topic is fragmentary, with different levels of coverage in different disciplines. For instance, the papers cited in this report appear in peer reviewed journals and reports in health, crime, social services, economics, education, and other sectors. Very few

involve cross sector citation. Economists usually cite only economists for example. The health people often cite only their brethren. The ACM Journal of Data and Information quality relies almost entirely on work information managers and computer scientists. And so on.

To complicate matters further, the administrative agencies that maintain records do not usually publish the results of their explorations of errors of any kind on the contents, processing, transfer, or correction of records. In this report, we rely mainly on peer-reviewed reports. Exploitation of dependable, but unpublished, reports generated solely for internal use by administrative agencies will have to come later.

Statistical texts on missing data and on non-sampling error in censuses and surveys are a readily available resource. In such texts, "non- response" in surveys is a cousin to "missingness" in administrative records. More distant cousins include texts on capture-recapture methods to estimate in the context of multiple data sources and elusive populations such as homeless or slave workers.

Most of the texts assume that the missingness is random, at times. They may assume further that missingness is random after variables that predict or are correlated with missingness are taken into account. Though virtually all such texts are written in a survey context, the basic ideas and methods are applicable in the context of administrative data systems. Any administrative system of records can be construed, for instance, as a census or survey of an identifiable target population. Little and Rubin's (2002) book is a useful source for handling technical issues, e.g. imputation of missing entries in records. We have been able to identify no statistical text on missingness that gets deeply to the *causes* of missingness or conceptual frameworks for missingness, however. But we identify useful empirical research in what follows.

"Measurement error" in statistical surveys is related to error in reporting, processing, or entering information in administrative records. See for instance, basic resources such as Lessler and Kalsbeck (1992), and Groves et al. (2009). We capitalize on the kinship relations in what follows and, moreover, try to make the limits on kinship explicit. For instance, books on survey methods and books on measurement error typically do *not* include the phrase "administrative records" in the list of contents or the subject index, nor do they dig into errors in this source.

Oskar Morgenstern's (1963, 1991) book *On the Accuracy of Economic Observations* set a remarkable precedent in attending to the quality of administrative records on which economists often rely. Wansbeek and Meijer (2000) and Manski (2007) advanced the state of the art in understanding the consequences of certain kinds of

error (omitted variables for instance) in analyzing data from administrative records and surveys. Informative papers have appeared at times in the economic literature, developed by Abowd and Vilhuber (2005) among others, for the case of employment and unemployment spells and flows. The rubric under which Abowd and Vilhuber put the matter is "sensitivity of estimates" of economic models, based on imperfect but improvable aggregates of administrative records.

Some peer-reviewed publications in applied statistics, evaluation, and social research, attend periodically to the quality of administrative records, though they do not usually do so in the context of IDS. The work often has implications for IDS and the IDS aims of producing dependable analyses. See, for instance, Joshi and Sorenson (2010) in the *Evaluation Review,* various issues of *Quality and Safety in Health Care and Medical Care*, and other journals cited in what follows. See also "Bridging the Gaps in Police Crime Data" (Maltz, 1999) at: http://bjs.ojp.usdoj.gov/content/pub/pdf/bgpcd.pdf. Reports from the U.S. Government Accountability Office, such as U.S. GAO (2010, 2012) are valuable and sometimes provocative in this context. The medical sector is one of the most substantial producers of peer-reviewed journal articles on the topic. This stems, at least partly, from efforts by the Institute of Medicine (IOM) to illuminate medical errors per se, and the record systems that support paid health care, e.g. the IOM's report *To Err is Human* (Kohn et al., 1999).

The resources include research on how to ask questions in the context of survey research, notably research falling under the rubric of cognitive sciences, e.g. Groves et al. (2009). We have uncovered no thoughtful applications of this line of work, however, in the context of administrative records, integrated or otherwise.

The relatively recent, pertinent, and general publications containing practical advice include Hatry's instructive chapter on "Using Agency Records" in the *Handbook of Practical Program Evaluation* (Wholey, et al., 2010), Herzog and colleagues' (2007) *Data Quality and Record Linkage Techniques*, and Groves et al.'s (2010) *Survey Methodology.* The last book is important for administrative records as well as surveys, though it does not consider administrative records deeply. The Herzog et al. (2007) book is especially interesting because it is relatively recent, more comprehensive, and more informative than its predecessors published between 1996-2001. These and other resources are relied on in what follows. The reader should, of course, exploit other resources when they appear. Keeping up with their appearances and understanding the products are not easy to do.

IDS DIMENSIONS, TAXONOMY, FRAMEWORKS

The definition of IDS given earlier is generic. To reiterate briefly, IDS involves putting together records from different sources on the same individual. IDSs can vary in kind and function. A crude taxonomy/conceptual framework for these kinds and their dimensionality can be constructed based on contemporary experience with IDSs. The crudity can be reduced with further work in the context of municipal, state, provincial, regional, national, or cross national records systems.

The time and temporal dimensions are important in any decent scientific framework for IDSs. An IDS, for instance, may be temporary. This is the case in many randomized controlled trials designed to estimate the effects of different social interventions. The relevant IDS may exist for less than five years. No integration beyond the period of the trial may be warranted if the framework for the science or the policy is so constrained.

Or, the time frame may be larger. An IDS that is supposed to inform understanding of the long-term effects of different ways of handling child abuse or neglect is likely to require ten-, twenty-, or thirty-year time frames. Or, the integration period may be indefinite. The point is that "integration," itself, has a time tag and duration. Failing to recognize each is unlikely to serve the public, or the statistical analyst, well.

Just as time is a fundamental dimension of an IDS, one can conceive of every element or entry in a subject's record as having a time tag, e.g. when the initial record entry was made. The out-datedness of record entries is a matter for quality control, pertains to privacy issues, and is important to statistical analyses. The updates on record entries also deserve time tags, e.g. when a correction is made or when a new event in the same variable category is entered. This topic is considered in a bit more detail later.

In more complex IDSs, administrative reports from more than two entities on the same individuals may be linked, i.e. integrated, at different temporal levels. For instance, some records may drive backward to (say) trace the ancestors of the individual whose records are of interest, and so involve retrospective reporting. Other systems may drive forward, to track progeny and track current events or behavior. Prospective longitudinal studies exemplify the latter genre.

A second fundamental IDS dimension is jurisdictional. Consider this first as involving integration of records at the same level of governmental jurisdiction, e.g. linking state records on an individual's reported income with state records on the same individual's state prison record. The jurisdiction may be national. One

federal exemplar is embodied by major efforts during the 1970s to link IRS records with census records, within the constraints permitted by law, to understand the relationships among reports of annual income made by the same people to each source. Despite the horizontal appearance of linkage in federal integration of this kind, one learns quickly that definitions, time tags, and the variable called income differ in character across agencies, for instance. They have to be accommodated. More about this is discussed below.

For IDSs, the challenge of integration across agencies that are hierarchical in character rather than horizontal is remarkable. In the United States, for instance, state education departments have had difficulty in obtaining records on students in schools, partly on account of differing interpretations of federal law, notably the Family Education Rights and Privacy Act (FERPA) regarding disclosure of the contents of students' records at the local versus state level. In addition, the definition of who is an enrolled student or the differences in the interpretation of the term "enrolled student" may differ appreciably from one school to another and may differ from the state's definition. Who is absent and who is not absent from a classroom on particular days, and who is a dropout, etc., has been controversial, at times, at each level. Hauser and Koenig (2011) and their colleagues have tried to bring more order to related dropout issues in their report for the National Academy of Sciences.

Issues Put into a Topical Checklist: A Taxonomy

Checklists are practical and often important. When reified at the institutional level, they become valuable, as in medicine, building construction, or aircraft safety, for instance. Gawande's (2009) *Checklist Manifesto* is a treat on this account. We put issues that need to be considered in a checklist here but do *not* yet put them in any priority order. This list could be construed as a rudimentary "taxonomy," if one wished to give more academic dignity to the idea. A taxonomy is conceptual, an essential part of any science, biological, physical, or social.

Coverage of the agency's target population, i.e. determining which individuals ought to have identifiable records and actually have them.

1. Definition of variables in the administrative record, their clarity and their uniformity across agencies and time tag.
2. Interpretations of definitions, by the subject of the record, by the recorder, and by others.

3. Jurisdictional variation in definition and interpretation of variables' definition.
4. Distortional influences that are random.
5. Distortional influences that are systematic.
6. Missing Data: Systematic and random missing information on particular variables in the record and on the complete record.
7. Local theories of distortion, missingness, and quality, including incentives and disincentives, time and burden.
8. Incentives.
9. Prospective and retrospective audit studies.
10. Processing.
11. Linkage.

Readers should recognize that some of the ingredients listed are those that concern the survey sample community. This is because administrative record systems can be thought of as a sample or census of some target population.

Record Quality Issues Endemic to IDSs, Explained and Illustrated

In what follows, the topics in this list are explained briefly. Examples are taken from peer reviewed publications and reports in different administrative contexts: police, social services, employment, health care and medicine, education, and others.

In creating and abiding by this listing, we depend on a theme in Rescher's (2007, page 8) book, *Error*:

"The fact of it is that knowledge can only advance across a battlefield strewn with eliminated errors."

Nothing is perfect.

COVERAGE OF THE AGENCIES' TARGET POPULATIONS

Any given mission agency has a target population whose members are eligible for services or who may be targets for the agency's attention and action. Each member of this population ought to have a record. Families who are eligible for and apply for food stamps are a matter of record, for instance. But not all who are eligible apply and become part of the record. Criminals who commit a felony are required to participate in the legal system. But not all criminals are caught. Then they are not part of a police arrest record. People who are required to pay taxes are defined under various statutes, but not all are identified by the jurisdiction's revenue service, for a time at least. So they do not become part of the record system.

For any administrative agency that creates records on its target, the gap between the target population on whom records ought to exist and the actual population (a sample of the target) on which records are created, is an indicator of record system's quality. Many "surveillance" systems, for instance, attempt to assure that, as all target individuals or events that should be detected are, indeed, detected and recorded. "Non-coverage" issues are treated in some texts on evaluation such as Rossi, Lipsey, and Freeman's (2004). Records of recipients of food stamps, for instance, can produce estimates of rates of the needy that differ appreciably from sample surveys of the eligible target population. Non-coverage is treated, at times, in studies oriented toward understanding completeness in coverage for particular kinds of statistical surveillance. Dichter and Rhodes (2009), for instance, provide empirical evidence that emergency room reports on interpersonal partner violence (IPV) can increase IPV detection by 30 percent over a rate based solely on police calls.

One broad approach to understanding lies in work by criminologists who compare the data generated from one administrative record system to surveys, or compare the data from one administrative record system to another.

In the United States, for example, both the FBI's Uniform Crime Reports (UCR) and the National Crime Victimization Survey (NCVS) aim to illuminate the incidence levels and trends in crime rates. The UCR depends on administrative records supplied to the FBI by police jurisdictions on the number of crimes they record and provide to the FBI for various types of crime. The UCR indexes seven crimes, homicide being one, in each police jurisdiction. The reports depend on each jurisdiction's crime definitions and procedures. The NCVS, on the other hand, depends on a national probability sample of people and elicits information on their victimization. Pepper, Petrie, and Sullivan (2010) provide a detailed review of such work. One finds, for instance, that in 2005, the UCR incidence for rape is at 0.3

per 1000 in the population and the NCVS is nearly double at 0.5 per 1000. The discrepancy is bigger for property crimes with UCR at 34.3 per 1000 and the NCVS at 154 per 1000. No direct linkage of records is *currently* possible at the individual subject level or at sub-national jurisdictional levels, as one might do in an IDS, given the differences in the character of the two sources. In principle, however, one can imagine a variety of linkages, integrations so to speak, at some sub-national level through small area imputation, for instance, if not at the individual level.

A second, and related example, of a way to understand the quality of records stems from work by Franklin Zimring (2011). A professor of law at Berkeley, Zimring depends on police records and is industriously suspicious about their accuracy in his efforts to understand the effects of policing strategies on crime. Police reports on homicide rates made to the UCR, for instance, may be biased downwards, to make the police department look effective in its work locally, or the reports may be biased upwards to justify hiring more police. Zimring checks the marginal frequencies reported to UCR against the marginal frequencies of medical examiner reports, another kind of administrative record system that one may assume is independent of police reports. Similarly, he checks statistical data submitted to UCR on auto thefts against related data available from the National Crime Insurance Bureau (http://www.ncib.org), which compiles records on payouts for auto theft. It appears that there is sufficient independence in the aims in each source to reckon that one is a reasonable check of the other's accuracy. Zimring acknowledges that such checks are not always possible, e.g. comparisons of two different record sources on the marginal rate of "assault."

A second, broad statistical approach to the coverage issue, and one that is more pertinent to IDS, lies in the literature on "capture-recapture" methods for statistically estimating the size of an elusive target population. Imagine, for instance, that one has two independent and presumptively random samples of a population of, say, prostitutes in a city. One sample is taken from health care provider records. Another is taken from police records. Assume that the subjects' identities in each record system are known and can be used as a basis for record integration. In the simplest scenario, an estimate of the total population size can be made using simple probability theory, knowing the number of people who have records in which they are identified by occupation as a prostitute in *both* record systems and knowing number of people identified as prostitutes in each independent record system. This magic depends on simple probability calculations and assumptions about the independence of the sources, among other things. In more advanced forms, the approach has been used in a variety of sectors when the sources of counts can be construed as independent or

any non-independence can be understood. See, for instance Lum, Price, Guberek, and Ball (2010) for a variation on the approach, important assumptions, and an application in the context of human rights violations in Colombia.

Evaluative research on the adequacy of coverage of administrative record surveillance systems is remarkable but choppy. For instance, one can uncover periodic work on the adequacy of post-marketing surveillance of drugs and medical device accidents, on "near misses" in airplane flights, on "homeless," crime and interpersonal violence, and vaccination rates in developing countries, among other topics. The choppiness of reports on the adequacy of administrative records in this context offers opportunity for learning more. The scenario begs questions: How and when can high quality "coverage" studies be mounted? What kinds of coverage studies? How do we accumulate knowledge and synthesize the research on results from such studies? How do we portray results and use the information in training and design of better surveillance systems and related systems?

For IDSs, the issues invite questions about whether and how "coverage" studies can be themselves integrated or made coherent across public health, social services, justice, and other sectors. And for capture-recapture related approaches, exploring empirical evidence on the independence of the sources of records, on the closed status of the population, linkage perfection, and technical matters such as homogeneity in the probability of detection (recording) in the systems seems important.

DEFINITIONS OF VARIABLES IN ADMINISTRATIVE RECORDS

A fundamental variable in all administrative record systems, integrated or not, is the identity of the individual on whom the record is maintained. The subject of the record, an individual, may identify himself or herself in the same way in each encounter with an agency, for instance, when a welfare check is issued. Or the individual may vary the identification of some encounters with the welfare agency and may differ in self-identification in an interaction with, say, a police agency, a court, or a hospital. The variations may be simple, such as leaving out middle name or initial in some encounters. The variation may lead to more extreme ones (for the middle and upper income folks) such as using an alias or an abbreviated name (Bob versus Robert). Any given person's recorded social security number may vary over time and across agency encounters. See, for instance, Abowd and Vilhuber (2005) on wage records in the context of unemployment statistical analysis.

Legal definitions of variables that characterize an individual's behavior can vary considerably from one jurisdiction to another. An assault with a knife may be classified as a misdemeanor in one police jurisdiction, for instance, and classified as a felony in another. This is despite the fact that the assault in each case had the same medical consequence. Similarly, being classified as "homeless" in an administrative record in Chicago does not have the same administrative/record meaning that it has in Miami. Definitions of abuse/neglect, high school dropout, autism spectrum, truancy, attendance, etc., all vary. Lawyers and lawmakers get paid to understand, unpack, and reify these things. Sometimes, they do well. But sometimes they do not do so, from the point of uniform scientific definitions.

The variable called "living or dead" may seem pretty clear. But administrative records are tied to place and time, and context counts at times. So, for example, the U.S. Veterans Administration's PROMISE system has elicited information on death of patients in end-of-life-treatment from Veterans Administration facilities with responsibility in this sector. Deaths in hospitals, for instance, were recorded well. But apparently, deaths in the hospice facilities to which veterans had been referred following treatment in hospitals and related facilities were not. This issue appears to have been corrected. False reports of death appear to have been far more rare, but important.

Time is a factor in all definitions. The variable called "poverty level" is one that receives the attention of economists because the definition can change, and probably should change, over time, as understanding of the idea increases. In a different area, the definition of unemployment, for instance, was targeted for political redefinition under President Nixon's administration. The relevant record-keeping agency, Bureau of Labor Statistics, had a tough time dodging a redefinition that served only to make the administration's employment policies look good.

The implication for IDSs is this: Changing definitions of variables in records have to be understood and, as a consequence, time tags ought to be routine for definitions of variables in record systems in which records are integrated. More to the point, if the leader of a statistical analysis of records in an IDS does not recognize time-related differences in definitions that are tied to different agencies that contribute to the IDS, the leader will be in deep yogurt.

Expert testimony about the challenges presented by records on income from different sources is not uncommon. Spar (2009), for example, declared that, "there is no way to analyze and reconcile the many measures of income between and within agencies. Each agency creates its own web site and its own data dissemination system with

little or no regard for the user, who has to go to more than a dozen sites and learn a dozen approaches to get a complete review of the socioeconomic data of the United States" (pages 21-22). The U.S. Government Accountability Office has taken a role in fostering order generally in federal systems. Some of the work, done in the interest of administrative interoperability rather than research, per se, directs attention to standards including standardized reporting across agencies, e.g. US GAO (2009).

Hatry (2010, p. 246) offered suggestions for alleviating time-related and interagency differences in definitions under the rubric of "Unknown, different, or changing definitions of data elements:"

- Make feasible adjustments to make data more comparable; define in detail each data element.
- Focus on percentage changes rather than absolute values.
- Drop analysis of data, such as data elements, when the problem is insurmountable.

This topic in our checklist might be handled in practical but more difficult ways through standardization of definitions across time, agency, jurisdiction, etc. The next section contains some related work. In related vernaculars, the word "harmonization" is used to get at the same idea.

INTERPRETATIONS OF DEFINITIONS

The person who is the subject of an administrative record may encounter a form asking about his or her race, and then may check one item in a checklist that declares that that he or she is white. The same person who fills in a related form encountered in another administrative context may declare she or he is black. Any such declaration may change over time. Recall, for example, changes in the way the U.S. Census Bureau and other federal agencies in the United States ask individuals about their race and ethnicity. Prewett's (2005) paper is instructive on this account. Similarly, the individual who completes an administrative record on the record's subject may, as a recorder, construe the record subject's race differently than another recorder with similar responsibility.

The correspondence between contents of administrative records on race/ethnicity and the self-reports of the individuals has been the subject of some interesting work. For instance, Kressin, Chang, Hendricks, and Kazis (2003) uncovered remarkable differences between entries in the patient treatment files of the Veterans affairs

records and the self-reports by the same individuals in an independent survey of their race/ethnicity and health status. Order of magnitude agreement rates were about 60 percent with lower rates for certain groups, notably Pacific Islanders, Asian, and Native American.

Australia's attempts to determine how many Aboriginal students there are in the education system have been frustrated, at times, by students who do not identify themselves in this category. Their interpretation and self-reporting may be on account of ethnic or cultural connotations or "self-identity" that precludes them using a particular descriptor for themselves.

The U.S. Census Bureau instituted a system for the 2010 U.S. Census for assuring that people with multiple ethnic/racial origins can "check all that apply." This helps to reduce naïve reliance on requirements that force an individual to choose only one descriptor, but does complicate the matter of classifying people. A respondent, for instance, could, in principle, choose from more than thirty different combinations to self-identify (Prewitt, 2005). Research on administrative systems that permit such choices is important for the future, but I have not been able identify any good examples thus far. Nor are we aware of research on systems that get to the matter of a choice of identity that the subject of the record finds threatening or socially undesirable.

One might expect certain administrative records to be accurate with regard to reporters' recording of gender and age of individuals. For instance, one might expect accuracy in reporting on these variables in a death certificate record. Rogers et al. (2004) explored the matter by linking records of people who were interviewed in a national health survey with records on the same individuals' later death certificates (death occurring between 1984 and 1990) in the National Death Index. They found that the administrative (certificate) records inflated age of deaths for males and deflated the age for females, and misreported race. The implication for some statistical analyses is that life expectancy can be overstated or understated for some groups. Further, estimates of correlations between variables, such as life expectancy and age, or education or income, may be inflated or deflated on account of the mistakes in records. This is a nontrivial matter if the country wants to understand its social capital, social security burden, and other matters.

One might expect that the idea and definition of student "success" in post-secondary education is clear. Colleges and universities are required to report this under federal law (Student Right to Know Act of 1990) and results are summarized in the Integrated Post-Secondary Education Data System. The system's Graduation

Rate Survey defines the rate as six-year graduation from the institution at which the student starts. Jones - White et al. (2010) examine the implications of redefining this. For instance, students who start at one institution then transfer to and graduate from another are excluded from the survey. Consequently, "success" may be underestimated. They propose using the Student Tracker (which covers transfers) of the National Student Clearinghouse instead, and use University of Minnesota, over three, entering freshman classes, as a case study. The difference between IPEDS rate (61 percent and the Track rate (71 percent) is substantial.

One might expect that the idea of interpersonal violence is defined uniformly, and is interpreted and recorded in uniform ways. See Joshi and Sorenson (2010) to learn about the issues, notably differences in the way the term is defined in police records and in the FBI's Uniform Crime Report (UCR), and in other reporting contexts, and frequent failure to record suspect-victim relationships.

The U.S. Government Accountability Office periodically assesses the quality of federal administrative records in certain sectors. A recent, and pertinent, such assessment concerns records used to evaluate the American Recovery and Reinvestment Act of 2009 (U.S. GAO 2009). Recipients of funding under the act report to the relevant federal agency. The reports vary notably in their interpretation of Full Time Equivalent (FTE) employment, despite early OMB guidance on the topic. Atypical and unexpected entries to records were not uncommon. The actions taken in response to the U.S. GAO assessments are given in other reports from the agency.

JURISDICTIONAL VARIATIONS IN DEFINITIONS AND INTERPRETATION OF VARIABLES

Variations in definition or in the interpretation of definitions of variables may occur at the individual level, as suggested earlier. The variation may be tangled with a local jurisdiction's rules, customs, or processes.

Events of the 1990s at Pennsylvania State University, and revealed publically during 2011-2012, are a case in point. They involved a football coach's alleged assaults on one or more children. Some controversy ensued on account of jurisdictional variation in definitions of sexual assault, rape, molestation, and other terms. Brisbane (2011, page 12) describes work by Wendy Murphy at the New England School of Law: "…in surveying the 50 states, she found something like 40 different terms to describe the act of rape of a child."

In a different context, the medical community, Bates et al. (2001) took on the task of understanding how to reduce frequency of errors in medicine partly through better use of information technology and a focus on coherence among various databases in medical jurisdictions. Their white paper emphasizes detection of error in a process made by a physician, for instance. But the message is subtle in recognizing errors generated by information technology. For example, different, but familiarly named, drugs may be misapplied, misspelled, or mis-recorded. The identities of two similarly named individuals may be confused. And so on.

The U.S Individuals with Disabilities Education Act (IDEA) has explicit definitions of disabilities that identify children whose special education needs are to be addressed under the act at http://idea.ed.gov. The federal definition provides minimum standards. States may augment these but not reduce them. The augmentation can engender state-to-state variation in records. MacFarlane and Kanaya (2009), for instance, examined such variation for the particular case of autism, using federal data and each state's legal codes defining the illness. Prevalence rates across states differed, mainly on account of the theme of Autism Spectrum Disorders (e.g. Asperger's Syndrome) which some states included by augmentation and others did not.

Criminal records depend on a local jurisdiction's definitions of particular crimes and on interpretations of these definitions by whoever is doing the reporting. In the United States, for instance, the definition of misdemeanor domestic violence in police reports may or may not cover sibling violence, violence against the elderly in a family, combat between gay partners, and other mayhem. These are apart from violence in spouse-like relationships. Matters become more complicated if the integrated data system is international. For instance, the European Union (EU) has deliberated about how to exchange information about criminal convictions in different EU countries. The idea of exchange is integration of a kind. Grijpink (2006) focused on chains of exchange and the definitional and other issues that emerged.

The notion of "graduation from college" seems plain enough. Nonetheless, different ways to get at graduation rates, based on institutions' administrative records, have engaged people's attention in the United States. For instance, the National Center for Education Statistics sends an annual survey to four-year institutions of higher education supported by federal Title IV funds, through the NCES Graduation Rate Survey. The more recent National Student Record Clearinghouse takes a different approach, targeting both Title IV recipients and institutions that do not receive such funding, and takes into account transfers from a junior/community two-year

college that have not graduated and that involve the student's enrollment in, and graduation from, a four-year institution. The rates of graduation, as one might expect, vary by state, and the order of magnitude of the average difference is about 6 percent. The issue is related to the "coverage" of target populations considered elsewhere in this report.

In the U.S., efforts to specify what a rural residence is and what is urban have a long history. The efforts are complicated by the fact that a rural sector can turn briskly into suburban and to urban in some countries and in regions within a country. The administrative service agencies that demarcate these have challenges even when the subjects' addresses are nominally correct in administrative record systems. For instance, Berke et al. (2009) examined how seven ways of classifying rurality could affect one's judgment about the use of U.S. Veterans Administration (VA) Services in rural areas of the United States. The notable variation in definition of rural leads to important differences in the counts of those who receive services and has consequently had implications for the hospital networks that provide veterans with services and for the VA's reporting of benefits to the Congress. Prospective studies of accuracy of this sort are discussed later.

DISTORTIONAL INFLUENCES THAT ARE RANDOM, MAYBE

Statisticians like *random* error in the contents of records, if there is to be any error at all. This is partly because recognizing random unreliability in observations and records of observations has a 400-year history extending from astronomy. It has a 200-year history in psychology and some of the other social sciences. In the contemporary context of administrative record processing, errors made in transcription, or in keystrokes, or in machine processing of machine readable forms, up to a point, might be assumed to be random, for instance.

This random error can be accommodated in many statistical analyses as long as one can assume that it is indeed random, and especially if one has an idea of its magnitude. Recall, for instance, the simple mathematics underlying "corrections for attenuation of a correlation coefficient" from elementary courses in educational and psychological testing. Or, the sophisticated user of fallible data might dump the measurement error into the category of random error in statistical or econometric models that treat the independent variables as "fixed" as opposed to "random."

Pepper, Petrie, and Sullivan (2010) reviewed the basic, classical random measurement error model in the context of criminological research, which depends heavily on

administrative records from police departments and related sources. The authors recognize, as others do, that the assumption of simple, independent random error makes it easy to understand the properties of statistical relationships and the parameters in more elaborate models, that indicate the strength of the relationship between the outcome, Y, and an assembly of predictor or explanatory variables, the X's on the right-hand side. These authors aver, more importantly, that a simple statistical model, which describes the report or response as a function of a true score and a random error, is often wrong and unpersuasive. Errors in administrative records on individuals (or in their self-reports in surveys) may, for instance, be related to the level of their crime or drug use, to their selective memory, to their selective responses to an interrogator, or to the interrogator's zeal in uncovering external corroborative or disconfirming evidence. See the section entitled "Distortions That are Systematic."

Adverse drug events (ADEs) and medical errors in health care settings are volatile topics in the developed economies such as the United States. One may then expect issues in record production and maintenance, and indeed they emerge regularly. Morimoto et al. (1996), for instance, studied the matter in health practices by (a) auditing practice records and charts, (b) surveying relevant physicians and patients, (c) surveying staff such as nurses and public safety officers, and (d) computer based triggers. The authors focused on the overlap in agreement among these sources. Overlap was startlingly low for inpatient records, with chart reviews uncovering 65 percent of ADEs and computer based triggers revealing 45 percent, and patients reporting at 4 percent. Outpatient surveys reported 92 percent of ADEs, and chart reviews identified 28 percent. Discrepancies of this sort invite thinking seriously about nonrandom error in some administrative records that contribute to an IDS.

DISTORTIONAL INFLUENCES THAT ARE SYSTEMATIC

At the local level, I may, as a police officer, arrest people often because I *like* to arrest people. I may, as a surgeon, prefer to do a particular kind of surgery, because I am good at it, it seems to work well, and it pays the rent. As a screener for children's admission to a Head Start Center, I may want to "help" a family get the child into the center by construing some of the family's income sources as irrelevant or by discouraging the applicant from reporting certain income sources.

My entry to the administrative record for which I have responsibility is accurate as far as it goes. It is I who made the entry in the record on an identifiable individual. I have had the discretion. This influence is important, but engenders no systematic

distortion of the contents of administrative records per se. Rather, the issue lies in the discretion that I, as the professional, have, and must have up to a point, in any surveillance or operating system. The administrative data systems that contribute to an IDS are often of this kind. The issue is related to the "coverage" issue identified earlier.

An individual who is the subject of an identifiable record used primarily for administrative purposes may have other reasons to systematically distort a report to a social worker, a cop, a judge, a physician or nurse, and so on. Systematic distortion may be a function of culture rather than intent. Asking some people from some countries in the eastern hemisphere about their age, for instance, inevitably involves some people who date their age from nine months before birth, as opposed to the western practice of giving age as calculated from the date of birth.

Pepper, Petrie, and Sullivan's (2010) review of measurement error in criminology attends seriously to systematic error as opposed to random error. One of the empirical investigations that they cite involves direct linkage between arrestees' reports of drug use and results of urinalyses on the same individuals. This was done to determine the quality of the reports that people give in an administrative (arrest) context. Roughly 15 to 30 percent of the individuals give inaccurate reports.

This kind of finding helps to justify the idea of periodic random audits of reports. It should also remind readers of the fallibility of laboratory tests, which then begs the question of how to assure better validation studies. Recall findings, for instance, on fraudulent, incompetent, or simply erroneous DNA testing entries in administrative records, fingerprint analyses, and so on in the forensic evidence research literature. A special issue of *Popular Mechanics* in December of 2009 covered the topic nicely for lay audiences. Pepper, Petrie, Sullivan's (2010) article provides interesting approaches to statistical modeling of the error in the contents of fallible records and illustrations of their application.

In the health sector, costs of services are a major issue. The accuracy of data or services and cost varies, for instance, among community mental health agencies. Wolff and Helminiak (1996) reported on processes in three such agencies. A major factor in the variation of data quality processes is, as one might expect, the fact that some are proprietary and some are not. The proprietary agencies appear to invest more resources in quality control when it is in their interest to do so.

Hatry (2010, p. 246) gives succinct advice on alleviating concerns about data accuracy. It is pertinent to systematic distortion and to random distortions:

- Check reasonableness of the data.
- Have someone check all or a sample of the data.
- Have a second person enter the same data and compare.
- Train or retrain the persons entering the data.
- Periodically audit at least a sample of the entered data against the original source.

"Reasonableness" checks are often defined operationally as examining the values of suspicious entries in the record. A female subject who is 14-years of age and whose record says that she has had four children should raise a question about the veracity of the reporting or the quality of the record entry e.g. Groves et al. (2009). Maltz (2010), for instance, advises us wisely, and in a practical way, to plot the data, rather than read through eye-glazing numerical columns and rows, so as to spot outliers, and he gives nice illustrations. Audits that are retrospective or prospective are a natural option for uncovering issues and potentially identifying what can be done about them, including training. Various approaches to audits are considered later.

We have not yet uncovered coherent and detailed approaches to "reasonableness" checks in the peer-reviewed literature on agency record systems, nor to the results of audits in the context of systematic distortion of record entries or information provided to the recorders. Much of the work on the topics is likely to be buried in internal reports of agencies and we do not yet have ready access to these or the resources to do so.

MISSING VARIABLES (OMITTED VARIABLES) AND MISSING DATA

The issue of variables being entirely missing from an assembly of variables that are essential to an analysis of what works, or of what is related to what, has a long history. Recall the following declaration from George Udny Yule in 1871:

> "Measurement does not necessarily mean progress. Failing the possibility of measuring that which you desire the lust for measurement … results in your measuring something else-and perhaps forgetting the difference-or in your ignoring some things because they cannot be measured."

With regard to variables that are not part of the record system, Dearden, Miranda, and Rabe-Heketh (2011), for instance, focused on the way statistical analysis of existing records can go wrong in a policy context. In particular, contemporary educational systems try to measure the "value added" of teachers and of schools in making administrative judgments. Teachers' schools are then rewarded or penalized based on the value-added analyses. Unbiased estimation of value added in either case depends on adjusting for background characteristics of children in the classrooms. This British study by Deardon et al. (2011) illustrated how the proxies for mother's education that appeared in administrative records were inadequate in adjusting for these variables, using mothers' education and other variables (Anderson, 2010). Recall earlier references to Manski and mis-specification of statistical models.

More generally, in the United States, important predictors of a child's performance, such as mother's education or household resources (income of family, books in the home), are absent from school-based administrative records and proxies from administrative records are used instead in estimating the effects of particular teachers, schools, types of teachers, or types of schools, and so on. This use of proxies may be in the context of school value-added analyses or teacher value-added analyses. Or, it may be in the context of estimating effects of intervention programs based on quasi-experimental designs that involve "adjusting" for differences between intervention groups and control groups. The typical recorded proxy is the child's or the school's participation in federally sponsored school lunch programs. The variable is used mainly because data for it are readily accessible and because obtaining more relevant and dependable data on family resources, such as income, mother's education, and so on, is costly. Of course, the bad decisions made on the basis of the proxy may be costly too, as John Easton (Director of the Institute for Education Sciences) pointed out in his remarks to the Intelligence for Social Policy conference in 2011.

The left-out variable problem for statistical analysts usually falls under the rubric of "omitted variables" or misspecification of the statistical model that is presumed to underlie the structure of the data. See, for instance, Charles Manski's (1995; 2007) work in econometrics. I have been able to uncover no serious attempt thus far, however, to connect the misspecification problem in statistical and econometric modeling to the presence or absence of variables in administrative record systems, or articles on what the implications of the misspecification might be for improving administrative record systems.

For a statistician who depends on a database, IDS or otherwise, "the best way of handling missing data is not to have it," said Lincoln Moses (personal communication

circa. 1985), a biostatistician given to brevity. Administrative records on a particular subject are, at times, entirely missing. The missingness may occur because of the gap between the target population and the actual population served, as discussed above. Some of the ingredients of the records, the entries on the variables, may be missing. For example, the missingness may be on account of errors of omission in processing a patient, client, or other subject of the record. Errors of omission occur, for instance, in medical contexts in which physicians may fail to check for and identify certain abnormalities or monitor drug therapy closely.

In the medical sector, Overhage et al. (1997), for instance, mounted randomized controlled trials to determine whether special software at physician teamwork stations could reduce such errors in the case of drugs. The approach depended mainly on creation and inclusion of "corollary orders" in workstation computers that reminded the physician team about what ought to be determined or checked once a particular drug regimen was prescribed ("trigger orders"). In effect, this trial is an early example of evidence for Gawande's *Checklist Manifesto* perspective, which advocates checklists, automated or not, for reducing medical omissions in action and record. The book, incidentally, does not consider error in the checklist *per se* or gaps in the checklist's entries.

Missingness in police reports that might be part of an IDS may not be trivial. Joshi and Sorenson (2010), for instance, reported that data on assaultive behavior were missing in about 40 percent of interpersonal violence reports, involving female victims, entered into police records in a large city. Data were missing for about 20 percent of incidents involving male victims in the reports. Gelman, Fagan, and Kiss (2007), for example, examined police reports on "mandated stops" and non-mandated ones in the interest of understanding whether and when full reports in the former differed from reports on the latter which did not require full reports. The aim was to learn about racial disparities in stops and reporting on the different kinds of stops. Though absolute values of rates changed depending on reports, the relative rates of stops involving different racial/ethnic groups did not.

The Veterans Administration's (VA) PROMISE system integrates records of data from surveys of families with administrative records provided by VA facilities in the interest of monitoring end-of-life care. A stereotypical form of missingness has involved reports from VA facilities about whether a particular patient died. Record of death may be sent belatedly to PROMISE by some VA facilities. To handle the problem, the requests for reports on health are sent twice, a month apart, for the same month. Assuring that people who are not dead are not missed in administrative

records is not as easy as one might think. A recent report of the Inspector General (IG) for the U.S. Office of Personnel Management (OPM) is a case in point (Office of the Inspector General OPM, 2011). The IG team discovered that about $600,000,000 over a five-year period had been paid out by the OPM to former federal employees who were retired or disabled, but who were dead during the same period. A son of a recipient of benefits, for instance, had been receiving payments to his dead father for 37 years. This was until the son himself died, and the mistake was discovered. One of the OPM approaches to resolving the problem has been annual matching of OPM records of payments to Social Security records of death. Another approach taken by OPM has been to sample recipients of payments, called "annuitants" in this vernacular, to determine who is dead and who is not.

The sampling strategy revealed that 6 of 4,400 annuitants were dead. There were 144 non-responses to the survey. This false positive rate of 3.4 percent or so may mean millions of dollars for the administrative agency. It can have substantial meaning also in statistical analyses of the effects of payout programs if the reduction in the cost of operations is of the same order of magnitude. At the sub-national level in the United States, Hatry (2010, p. 246) offers practical advice to manage the problem of missing or incomplete data in agency records in the context of evaluating program effects or operations:

- Go back to the records and related data sources (interview program staff, for example) to fill in as many gaps as possible.
- Exclude missing data, or provide a best estimate of the missing values.
- Determine whether part or all of the evaluation study needs to be modified or terminated.

References to statistical methods of imputing missing data were given above. We have encountered no approaches to understanding missingness at the local level administrative record systems that approach those used and described above for federal agencies.

In reporting on a statistical analysis of such data, telling readers how the missing data were handled is a virtue. This virtue is often not exercised. And the exercise varies in character. Published reports on analysis of complex data files in the education sector, for instance, vary in reporting handling from "…we used only complete data …," a practice that is vulnerable for obvious reasons, through simplistic imputation (substituting mean levels, etc.), out to ferociously complicated imputation methods.

Peugh and Enders, among others, (2004) in education research, report on how researchers handle missingness in analyses. See also the section on Prospective and Retrospective Audit Studies for approaches to understanding missingness.

LOCAL THEORIES OF DISTORTION, MISSINGNESS, AND QUALITY OF RECORDS IN AN IDS.

Partly because definitions of variables in a record vary across jurisdiction (misdemeanor crime for instance), the character of distortion or error in administrative records is localized. Moreover, the incentives for distorting, deliberately or indeliberately, may be local. Records of "suicides," for instance, may vary in character on account of local culture and religion, changes in leadership in the public health sector, and so on. This localization presents challenges to those who merely assume that suicide, for instance, means the same thing in a death record in Chicago as it does in Miami, Philadelphia, or Des Moines.

Variation in quality of records across agencies or sites can be considerable. Abowd and Vihuber's (2004) review of job history records, for instance, reports that "A Survey of fifty-three state employment (job) security agencies in 1996 to 1997 found that most errors are due to coding errors by employers, but when errors were attributable to state agencies, data entry was the culprit." They cite a Bureau of Labor Statistics (1997) study in support. More to the point, they extol the virtues of California's optical character recognition (OCR) system, as opposed to manual transcription. Other jurisdictions, of course, may not have the resources to employ OCR.

For statistical and social sciences at the aggregate level, the potential systematic distortions were recognized over thirty years ago, by Donald T. Campbell (1975, page 35). He declared:

> "I have come to the following pessimistic laws. The more any quantitative social indicators are used for social decision-making, the more subjective it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

We have, thus far, been able to uncover no serious empirical research on the extent to which Campbell's small theory can be sustained by evidence that is other than anecdotal. Such research would, of course, have to define and measure properly the use of the indicator and the conditions under which the use is one that is important enough to engender corruption. This is a big gap in the social sciences. It is relevant to the next topic.

PROVENANCE AND INCENTIVES

Criminologist Michael Maltz (2010, page 25) reminds his readers of a pertinent quote from J. Stamp, an economist writing in the 1920s:

> "The government are (sic) 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 in the first place from the *chowty dar* (village watchman) who just puts down what he damn pleases."

Put aside the fact that Sir Joshua Stamp got the phrase wrong in 1929 on page 258 of his tract; the right word is *chowkidaar* (according to my Indian colleagues).

For administrative system quality, the important implication is this: the provenance of the record's entries ought to be recorded. Who indeed supplied the information, to whom and when? In fact, provenance has become a basis for understanding quality in the context of management information and computer science guidelines. See, for example, articles in the *ACM Journal of Data and Information Quality.*

Further, in the arena of surveys and censuses, the U.S. Census Bureau has done continuous research on variation produced by different census takers in census records, among other important topics. This work is sometimes based on random allocation of census workers to different jurisdictions and analyzing components of variance attributable to census taker differences. Something akin to this is done in inter-laboratory testing to understand dependability of different laboratory reports on identical samples of blood, variations in interpretations of fingerprint samples, DNA samples, etc.

The matter of who is reporting is tangled with incentives to report, of course. Why people or institutions do not record accurately, or do not report at all, is a big topic. Michael Maltz's *Bridging Gaps in Police Crime Data* considers plausible reasons why police don't submit crime data to the FBI for the Uniform Crime Reports (UCR). There are a more than a few.

In another arena, McCarthy et al. (2000) examined accuracy of administrative records on coded diagnoses of medically discharged patients relative to retrospective medical chart review in nearly seventy hospitals in two states in 1994. Administrative records matched chart reviews most often (80 percent and up) for events such as reopening of a surgical site and post-operation heart attacks. The lowest match

rates occurred for "no evidence present" (9 percent) and post-operative wound infection (37 percent). Possible reasons for variation were examined and include: reimbursement systems, level of evidence used by the particular hospital team to code a particular problem as such, and other reasons.

Whitt (2006) takes seriously the idea that "suicide" is a social construction, subject to religious or other value laden interpretations, in examining New York City's suicide rates. He found a sharp decline in reported rates from 1985 to 1989 following the appointment of a particular medical examiner called A. The appointment of a newer medical examiner, call him or her B, led to a second shift that was more in accord with earlier records. Whitt reckoned that the sharp decline under A's regime was a matter of official interpretation and increased classification of certain deaths as being due to "other accidents." In a prospective audit discussed below, the U.S. Government Accountability Office found fraudulent representations of family income by Head Start screening staff. By disregarding certain income, for example, the staffer could enhance the likelihood that a child in the family would be admitted to the Head Start Program. The order of magnitude in misrepresentation was sufficient to justify concern and news coverage. But it is arguably small in relation to the magnitude of misrepresentations of assets in the banking and loan industry during the 2008 financial crisis and thereafter.

The foregoing examples are merely illustrative. I have uncovered no systematic reviews of evidence on the nature of incentives to distort and the magnitude of their effect. Exceptions include Maltz's citations of newspaper articles that describe systematic downgrading of crime counts in "Bridging Gaps" report. It seems sensible, regardless, to speculate about how the matter of incentives might be dimensionalized. The consequences of an accurate entry in a record, for example, may be negative for either the record's subject, as in the example of Head Start staff. Or the consequences may be negative for the recorder or the record agency, e.g. "open" police cases. A second dimension might be the burden of reporting or backtracking to determine the relation between earlier reports and later ones, e.g. medical charts versus administrative records in patient discharges. A third might be type of agency, e.g. judicial/enforcement versus ameliorative. It seems sensible to consider deeper systematic reviews of evidence on the topic if we want to learn more in the IDS context.

PROSPECTIVE AND RETROSPECTIVE AUDIT STUDIES OF DISTORTION AND MISSINGNESS

Most studies of the quality of administrative records are retrospective in character. For instance, a record whose entry is made on a given date may be checked for accuracy by re-interviewing the individual on whom the record was generated at a later date. The record's contents might also be checked against a second independent source of information or against earlier reports as in reviews of hospital discharge summaries against earlier patient charts (Callen et al., 2010).

Prospective studies of the quality of administrative records are uncommon. One prospective approach involves the creation and use of counterfeit clients or patients, or other targets of a potential record. These individuals are trained as actors to present themselves in a certain standardized way and in a particular administrative context. Records on these actors are generated (i.e. variables are coded) by the administrative agency's recorder who is unaware that the individual is a counterfeit. The results of this coding and other processing are then examined for accuracy. Consider the following examples.

To determine the accuracy of administrative records in selected health care systems, Peabody et al. (2004) used trained standardized patients (actors) who made unannounced visits to three health care sites. The benchmark for judging quality of the resultant administrative record was the standardized patient. Medical records on these standardized patients were then later examined to identify errors of commission and omission. The authors found that a correct primary diagnosis was given in 57 percent of records on these visits, with 13 percent of errors attributable to physician misdiagnosis, 8 percent attributable to missing encounter forms, and 22 percent to data entry errors.

The U.S. Government Accountability Office (2010) mounted a prospective study that involved people posing as needy in their applications for Head Start services for their children. *The Washington Post's* Nick Anderson reported results in the May 19, 2010, issue. "In eight of 15 cases, The GAO found, staff members at the centers fraudulently misrepresented financial information from applicants." The full report on methods and results is in U.S. GAO (2010); see http://usgao.gov. The GAO effort involved fifteen counterfeit scenarios deployed in each Head Start center in six states and Washington D.C. It appears from the GAO report that in all states and in D.C. that eight of fifteen scenarios were routinely falsified in the Head Start center's record system.

Retrospective audits are a far more common approach than prospective studies to understand missingness, distortion, errors, and so on. The idea of "verbal autopsies" received some recent attention in the health sector. The United States, Canada, and India have cooperated in a large-scale study to learn whether over 100,000 potential victims of malaria had actually had malaria from 2001 to 2003. Purported victims' families were surveyed to elicit information on the victims and their state prior to death. The reports were then sent to physicians who made diagnoses based on the information elicited about symptoms. The authors maintain that, "malaria can be hard to diagnose with only verbal autopsies." But based on the evidence that the authors uncovered, they declared that the death rate from malaria could be eight times the number recorded in agency records (World Health Organization, 2007).

PROCESSING OF RECORDS AND QUALITY CONTROL IN THE IDS

For serious researchers, evaluators, and information scientists, record processing is not a trivial matter. However, the topic is vast. For instance, a record whose entries are made by one entity or person at one time may then be sampled and edited by a second entity or person at another time. The edited record is the record that might then be used in statistical analyses. The editing is a part of processing, but if the editing process itself is suspect or insufficient, then problems remain.

Consider, for a moment, that a controlled randomized field trial (RCT) can be construed as an IDS, at least for the trial's duration. Consider further a specific controlled trial on an anti-violence intervention in Philadelphia, which involves screening of high-risk offenders who would then be identified as such and randomly allocated to different interventions. From Jordan Hyatt (2011), who is coordinating such a trial, we have the following:

> The challenges to the integrity of the random assignment process were largely a function of the characteristics of record-keeping within the First Judicial District (FJD). Screening for RCT eligibility, risk assessment, and unit assignments took place within APPD's Intake Unit located in the FJD's courthouse. Once a defendant was sentenced to probation, they were immediately instructed to report to the intake unit. The case file, including paper copies of the judicial order and docket information, remained with the courthouse clerk. Therefore, when the defendant arrived for screening, the intake clerks could rely only on the paperwork the defendant had in hand when creating the electronic file for that supervision case. This often resulted in

> the creation of a file without a record of a specialized condition (which would have eliminated the offender from the trial) but with the appearance of RCT eligibility. In addition to the incorrect unit assignment, the electronic record, relied upon by APPD for internal evaluations, would not reflect the court-mandated assignment until the error was discovered and resolved. Any discrepancies within the multiple sets of administrative records often remained undetected until the physical hard copy of the defendant's file arrived at APPD and was routed to the supervising officer … this resulted in probationers remaining in the inappropriate unit for extended periods of time, in addition to the complications this system posed for random assignment.

The foregoing reminds us that time tags are essential for understanding record ingredients and quality in a complex inchoate environment.

Mode of processing, notably method of transcription, is important for many beyond the temporal flow. For instance, Callen, McIntosh, and Li (2010) focused on the accuracy of the entries on medication in hospital discharge summaries that were handwritten as opposed to electronic. The standard was retrospective examination of inpatient records, i.e. patient charts. Reviewing nearly 2000 records of and the related charts, the authors found 12-13% error rates in records on medication regardless of the level of training of the physicians involved. Cardiovascular drugs appear to have been most commonly omitted or misreported in the summaries.

LINKAGE

This report focuses on quality of records in the data systems that are integrated rather than on integration, i.e. record linkage, *per se.* The quality of linkage among records is itself a matter for concern. The metrics for judging quality include, among others, the recall rate (proportion of true matches that are designated as such) and false non-match rate (true matches that are designated as non-matches). These can be put into a Bayesian context to understand the probabilities involved and the conditional probabilities. Herzog et al. (2007) give the basic mathematical notation, equations, and explanation.

Suffice it to say that the quality of matches, the linkages, depend on the quality of the coverage of each source of records, the quality of entries to the subject's records in each, the missingness, and other factors described earlier. The methods

of linkage under various assumptions about completeness of records or accuracy of their ingredients are given by Herzog et al. (2007) in a frequentist context, Bayesian context, and algorithmic decision rules that led to different levels and kinds of matches. The book is instructive in these ways, but it does not handle deeply the record quality issues that we have identified above or the research on them.

Standards and Standardization

The National Center for Education Statistics (NCES) has initiated Common Data Standard Initiatives to develop voluntary, common data standards for a "key subset of K-12 … and K-12 to post-secondary education transition variables" (http://www.commondatastandards.org). The subset includes demographics, program participation, and course information. An ambitious effort, the aim is to secure voluntary commitment to correction and use of uniform information. The category of student identification, for instance, includes thirteen ingredients, including first and middle name, surname, maternal last name, last/surname at birth, nickname, alias, etc. This particular class of data elements is also subject to "special designation" on privacy grounds (Personal Individual Identifiers (PII)) under various federal privacy laws.

The Family Education Rights and Privacy Act (FERPA) and associated regulations (34CFRpar99) acknowledge the "right of the parent to inspect his or her child's record and, in the case of eligible students, the student's right to inspect his or her own education record for accuracy …" (SLDS Tech Brief, Brief 2, pages 6, 16). Further, the laws are construed to encourage or require data managers to develop procedures to assure that record entries are up to date, complete, and accurate. In particular, periodic audits of data quality should be undertaken by either substantiating the quality of the individual data elements or identifying inaccuracies for correction. Periodic quality audits should be built into data collection, reporting, and release cycle (SLDS, Page 16). This includes "describing steps required to validate data … data entry, checking for errors, confirming that errors are real and not outliers, identifying each place the incorrect data element is stored in the data system, and providing corrections to the data entry staff" (Page 16).

The U.S. Office of Management and Budget (OMB) (1980) has standards for exact matching and statistical matching that are pertinent here. *Exact matching* refers to the use of a unique identifier of a subject to link the subject's record from one source with the same subject's record from another source. The phrase *statistical*

matching refers to scenarios in which no unique identifier is available and the aim is to construct similar groups in the interest of (say) estimating the effects of interventions. The issues considered in regard to accuracy and missingness of records or record entries are pertinent in principle to both exact matching integration based on clear identifiers and to statistical matching. However, we have not had the time to explore the relevance or practicality of this OMB work to current concerns. The advances in statistical matching over the past twenty years, however, suggest that the OMB standards will be out of date in this respect.

Standards in the statistical, social, educational, and other sciences often are conscientious about definitions of terms within the particular standard set. Examples are above. However, for an IDSs, the definitions of terms may differ across the agencies that contribute records to the IDS. Herzog et al. (2007) enumerate some practical approaches to harmonizing across record systems. Their counsel includes: Standardization of Terms; Standardization of Spellings; Consistency of Coding; Elimination of Entries that are out of Scope; and Integrity Checks. Each form of standardization presents challenges, of course, when undertaken across agencies that each have their own mission, vernacular, culture, and power.

How Much Record Quality for What Purpose and at What Cost?

The topic inherent in the question is related to the concept of "fitness for use," which has been offered as a criterion in judging the quality of administrative records. It is related to the idea of "consequential validity" in the area of academic achievement testing offered by Samuel Messick, among others. The notion hinges on the fact that the record system records must be appropriate for the use at hand. To take a simplistic example, records of the weights of individuals based on their having been measured using a truck stop scale are unlikely to satisfy either the statistician interested in averages or the clinician interested in the person at hand.

Medieval Jewish history, medieval Arabic history, and early U.S. colonial history, provides some interesting lessons. We retreat here to history because we have been able to uncover little work, as yet, on the costs of errors and the relative benefit/cost ratios attached to various schemes for improving the quality of records, whether the records are integrated or otherwise.

In the fourteenth century, for example, rabbis argued about how to estimate the value of an olive crop in the interest of tithing to the Temple. Should one hire a trustworthy, but low-wage, assistant to go to the vineyard to grab a sample of olives, and declare the value of the crop based on this inexpensive information source? Or, should one develop a plan for a scientific sample or full-blown administrative system, issue a Request for Proposal, vet the competitive bids, and so on, in order to determine the crop's value? One method is quick and cheap, but results may be dirty. The alternatives are slow and expensive, but cleaner and more transparent.

The rabbis deliberated for a period of time. That was what the rabbis were paid to do. They created a decision function. In particular, they declared that if the information was necessary only for bureaucratic purposes, one could rely on the quick and dirty approach. If, on the other hand, the demand for the information had a higher purpose, if it was demand from the Deity, then one ought to expend serious resources to get the accurate estimate of value. Where the U.S. Congress or state legislatures, or county officials stand, in relation to the Deity, is not clear. But, the decision rule seems sensible. That is, the resources invested in generating the information ought to be in accord with the origins of the demand and the information's expected uses.

Put into somewhat different terms, one might then be satisfied with some records despite their imperfections, if the consequences of making a decision based on this statistical data are not serious in the decision maker's view (the origin of the request). If the consequences of a wrong conclusion based on imperfect records are serious, then substantial investments in assuring record quality are warranted.

In medieval Arabic literature, Ibn Khaldun inveighed evenhandedly against Christians and Muslims for inflating body counts, an indicator of who won or lost the battles in the thirteenth century and earlier. This stereotypic form of corruption of records has a rich history to judge from Phillip Knightley's *"The First Casualty: The War Correspondent as Hero and Myth-Maker from the Crimea to Iraq."* The book traces the problem from the Napoleonic Wars up through Iraq (Knightley, 2004). Such records are vulnerable for reasons given earlier in the context of Don Campbell's (1975) remarks on social indicators: once the thing becomes a basis for political or financial decisions, it will be corrupted, at least up to a point.

In the matter of estimating costs of errors in administrative records, we have been able to locate a few disaster stories but nothing in the way of serious systematic reports in the peer-reviewed scientific literature. One such story, covered widely in the popular press in New Jersey, concerned the state bid for substantial federal grant

moneys in the "Race to the Top" competition (Friedman, 2010). The state failed to achieve a high enough rating of its proposal to get funding. An error had been made in the New Jersey bid by using the N.J. budget for one fiscal time frame instead of the budget figures for the prescribed time frame. The state was deprived of millions of dollars. The state's governor blamed the feds for being picky. The feds exercised rectitude.

In the matter of estimating the benefits of improved quality or linkage in monetary terms, we have been able to uncover no detailed handling of the topic or dependable numbers in the published literature. Herzog et al. (2007) however gives succinct examples of how private corporations have enhanced quality and actualized integration (which they label as linkage) to advance the organizations' aims. The descriptions explain how, for instance, competitive advantages were achieved in this way by Harrah's hotel and casino chain, Choice Hotel International, Wal-Mart, Federal Express, and Albertson's pharmaceutical and prescription services. The Fed Ex illustration, for instance, is explicit in attending to the organization's support and use of new software to properly register and unify handling of addresses, signature requirements, and other details that appear in bills but were often not sufficiently coherent and uniform to permit linkage.

Dual Mission Agencies

It is difficult to find published reports on the consequences of making decisions based on imperfect quality in administrative records when the records are used for statistical analysis. Publications by economists may be an exception. Abowd and Vilhuber (2004), for instance, give order of magnitude estimates to the mistakes that one might make based on analyses of imperfectly identified wage earners and the unemployed. The bias in estimates of interrupted job spells for instance, can be around 11 percent once the records are corrected. For the decision maker at the local level, such an error could mean bankruptcy. At the level of the federal government, the error means being off by well over $100 million.

Dual mission agencies here include the U. S. Internal Revenue Service in which responsibility lies with getting people to pay the "proper" amount of taxes" and in reporting statistics on income. They include police departments whose responsibility lies in enforcing law and perhaps in preventing crime, and also in reporting statistics on crime. The agencies in the United States include the FAA, which has a surveillance system for "near misses" as well as an enforcement mission. In each,

there is a remarkable tension between the need to identify specific cases or incidents and to report statistics on the incidents. The incentive to report well on incidents is arguably different from the incentives to report on incidence. I have not had the time to reconnoiter the research that these agencies have done so as to understand the quality of their administrative records whether integrated with others or not.

Concluding Remarks, but No Conclusions

Understanding how to arrange one's thinking about the quality of administrative records, whether integrated or otherwise, is no easy matter. How indeed should one, or can one, assess quality and its dimensions, how to handle variations in definitions and their interpretations, time tags for currency of entries to records, and what the implications for analysis in different contexts are?

Work on the topic cuts across different disciplines and organizations, across differing levels of governmental and nongovernmental agencies. It cuts across the varying functions of the records, in vertical organizations and horizontal ones. The checklist approach, stressed here, may seem pedestrian to some readers. To elevate its import, think about "taxonomy," a good and more dignified word to get at similar ideas in the sciences.

In brief, look at: coverage of the ostensible target population; definitions and the interpretations of them; variations in definitions across jurisdiction and disciple; provenance, time tags, and continuity of measurement or recording; and systematic error and random error. And look at the local empirical characterizations of each in the systems to which the IDS is directed. Inventing local theories about the nature and magnitude of distortion and other error in records seems a good thing to try to do. The quality of the integration and the product depends on each ingredient in the checklist/taxonomy.

The idea of an IDS is big. The questions and ideas that an IDS initiative provokes are also big. Whether and how to harmonize definitions is a fine challenge. How to standardize measurement or observation, through training or in other ways, seems very important. How to understand when and how the investments in harmonization or other improvements are warranted, and how they might be actualized, is a ferocious challenge.

It will be a delight to see what younger colleagues do.

References

Abowd, J.M. & Vilhuber, L. (2005) The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers. Journal of Business & Economic Statistics, 23(2): 133-152.

Anderson, N. (2010, May 19). Officials find evidence of fraud at Head Start centers. The Washington Post. http://washingtonpost.com/wpdyn/content/article/2010/05/18/AR2010051803400.htmwpisrc=nl_fed

Berke, E. M., West, A. N., Wallace, M., E., & Weeks, W. B. (2009) Practical and Policy Implications of Using Different Rural-Urban Classification Systems: A Case Study of Inpatient Service Utilization among Veterans Administration Users. *Journal of Rural Health*. 25(3), 259-256. DOI:10.1111/j.1748-0361.200900228.x

Boruch, R. F. (1984) Ideas about Social Research, Evaluation, and Statistics in Medieval Arabic Literature: Ibn Khaldun and al Biruni. *Evaluation Review*. 8 (6), 823-842.

Boskin, M. J., Dulberger, E., Gordon, R. J., Griliches, Z., & Jorgenson, D. W. (1998) Consumer Prices, the Consumer Price Index, and the Cost of Living. *Journal of Economic Perspectives* 12(1), 3-26.

Brisbane, Arthur S. (2011) Confusing Sex and Rape. *New York Times, Sunday Review* (November 20 2011), page 12.

Callen, J., McIntosh, J., and Li, J. (2010) Accuracy of Medical Documentation in Hospital Discharge Summaries: A Retrospective Analysis of Medication Transcription Errors in Manual and Electronic Discharge Summaries. *International Journal of Medical Informatics*, 79(1), 58-64. Doi:10.1016/j.jmedinf.2009.09.002.

Campbell, D.T. (1975) Assessing the Impact of Planned Social Change. In Gene M. Lyons (Ed) *Social Research and Public Policies*. Hanover, NH: Public Affairs Center Dartmouth College, (page 35 Quote), pp.3 – 45.

Dearden, L., Miranda, A., and Rabe-Hesketh, S. (2011) Measuring School Value Added with Administrative Data: The Problem of Missing Variables. *Fiscal Studies*, 32(2), 263-278. doi:10.1111/j.1475.2011.00136.x.

Dichter, M. and Rhodes, K. V. (2009) Reports of Police Calls for Service as a Risk Indicator of Intimate Partner Violence. *Academic Emergency Medicine*, 16, 83-86.

Florence, Curtis, Shepard, Jonathan, Brennan, Iain, and Simon, Thomas (2011) Effectiveness of Anonymised Information Sharing and Use in Health Research, Police, and Local Government Partnership for Preventing Violence Related Injury: Experimental Study and Time Series Analysis. BMJ2011:342:d3313.doi10.1136/bmj.3313

Friedman, M. (2010, August 25). Governor Christie Blames Washington Bureaucracy for State's Failed "Race to the Top" Application. *New Jersey Real-Time News*. Retrieved from http://www.nj.com/news/index.ssf/2010/08/gov_christie_blames_washington.html.

Gawande, Atul (2009) *The Checklist Manifesto: How to Get Things Right.* New York: Metropolitan Books.

Gelman, A., Fagan, J. and Kiss, A. (2007) An Analysis of New York City Police Department's "Stop and Frisk" Policy in the Context of Claims of Racial Bias. *Journal of the American Statistical Association*, 102 (479), 813-823. Doi:10.1198/016214506000001040.

Grijpink, J. (2006) Criminal Records in the European Union: The Challenge of Large Scale Information Exchange. *European Journal of Crime, Criminal Law, and Criminal Justice* 14, 1.

Groves R., Fowler F., Couper M., Lepkowski J., Singer E., & Tourangeau R. (2009) Survey Methodology. New York: Wiley.

Hatry, H. P. (2010) Using Agency Records. Chapter 11 of J. S. Wholey, H. P. Hatry, and K. E. Newcomer(Eds) *Handbook of Practical Program Evaluation*. Third Edition. New York: Jossey-Bass/Wiley, pp. 243-261.

Hauser, R. M. and Koenig, J. A. (Eds) (2011) *High School Dropout, Graduation, and Completion Rates: Better Data, Better Measures, Better Decisions*. Washington DC: National Academies Press.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007) *Data Quality and Record Linkage Techniques*. New York: Springer.

Huxley, T.H., 1893. Evolutionand Ethics. Collected Essays. Macmillan, London.

Hyatt, Jordan (2011) JLCG Internal Report, Philadelphia Anti-Violence Experiment. October 14, 2011. Jerry Lee Center for Research I Criminology, University of Pennsylvania, Philadelphia Pennsylvania.

Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Institute of Education Sciences, U.S. Education Department (2010) SLDS Technical Brief/ Guidance for Statewide Longitudinal Data Systems (SLDS) *Data Stewardship Managing Personally Identifiable Information in Electronic Student Records*. Washington DC: IES USDE (NCES 2011-602).

James, Henry. (1907). Preface. In Roderick Hudson. New York: Charles Scribner's Sons http://www.henryjames.org.uk/prefaces/text01.htm.

Janson, Carl-Gunnar (2000) (Ed) S*even Swedish Longitudinal Studies*. Stockholm Sweden: Swedish Council for Planning and Coordination of Research

Jones-White, D.R., Radcliffe, P.M., Huesman, R.L., & Kellogg, J.P. (2010). Redefining student success: Applying different multinomial regression techniques for the study of student graduation across institutions of higher education. *Research in Higher Education*, 51(2), 154-174.

Joshi, M. and Sorenson, S. (2010) Intimate Partner Violence at the Scene: Incident Characteristics and Implications for Public Health Surveillance. *Evaluation Review*, 34 (2), 116-136.

Knightley, P. (2004). The First Casualty: The War Correspondent as Hero and Myth-Maker from the Crimea to Iraq. Maryland: Johns Hopkins University Press

Kressin, N. R., Chang, B. H., Hendrilcks, A., and Kazis, L.E. (2003) Agreement between Administrative Data and Patients' Self Reports of Race/Ethnicity. *American Journal of Public Health*, 93(10)1734-1739. Doi:102105/ajph.93.10.1734.

Kohn, L.T. and others (1999) *To Err is Human: Building a Safer Health Care System*. Washington D.C. Institute of Medicine.

Lessler, J. and Kalsbeck, W. (1992) *Non Sampling Errors in Surveys*, New York: John Wiley and Sons (Wiley Inter. Science).

Little, R. and Rubin, D. (2002) *Statistical Analysis with Missing Data*. (Second Edition) New York: John Wiley and Sons (Wiley Interscience)

Ludwig, J. and others (2011) Neighborhoods, Obesity, and Diabetes--A Randomized Social Experiment. *New England Journal of Medicine*, 365, 1509-1519.

Lum, K., Price, M., Guberek, T., and Ball, P. (2010) Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007. *Statistics, Politics, and Policy*, 1 (1), Article 2. DOI10.2202/2151-7509.1005. Available at: http://www.stat.duke.edu/~kcl12/Human%20Rights_files/academic-paper.pdf

MacFarlane, J.M. & Kanaya, T. (2009). What Does it Mean to be Autistic? Inter-state Variation in Special Education Criteria for Autism Services. J*ournal of Child and Family Studies*, 18.

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality, 1(1), 1-22.

Maltz, Michael (1996) Criminology in Space and Time: Life Course Analysis and Micro-ecology of Crime. In John Eck and David Weisburd (Eds) *Crime and Place*. Monsey New York: Criminal Justice Press.

Maltz, M. (1999). Bridging gaps in police crime data. Bureau of Justice Statistics. Washington, DC: U.S. Department of Justice.

Maltz, Michael (2010) Look Before you Analyze: Visualizing Data in Criminal Justice. Chapter 3 of: A. Piquero and D. Weisburd (Eds) *Handbook of Quantitative Criminology*. New York: Springer, pp 25-52.

Manski, C. F. (1995) *Identification Problems in the Social Sciences*. Cambridge MA: Harvard University Press.

Manski, C. (2007) *Identification for Prediction and Decision*. Cambridge MA: Harvard University Press.

McCarthy, E.P., Lezzoni, L.L., Davis, R.B., et al. (2000) Does Clinical Evidence Support ICD-9-CM Diagnosis Coding of Complications? *Medical Care*. 38(8); pp. 868-876.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). Washington, DC: American Council on Education & National Council on Measurement in Education.

Morgenstern, O. (1963/1991) *On the Accuracy of Economic Observations* (Second Edition).

Morimoto, A.M., Jordan, K.C., Tietze, K., Britton, J.S., O'Neill, E.M., & Ruohola-Baker, H. (1996). Pointed, an ETS domain transcription factor, negatively regulates the EGF receptor pathway in Drosophila oogenesis. Development, 122(12), 3745-3754.

Morimoto T., Gandhi T.K., Seger A.C., & Bates D.W. (2004) *Quality and Safety in Health Care:* 13:306-314.

Murray, Christopher J. L. and others (2012) Global Malaria Mortality between 1980 and 2010: A Systematic Analysis. *Lancet*, 379 (9814) 413-431.

National Center for Educational Statistics (NCES) (2010) *Common Data Standard Initiatives*. http://www.commondatastandards.org; http://www.ies/whatsnew (retrieved 5/24/2010.)

National Center for Educational Statistics (NCES) (2008) Graduation Rate Survey.

Office of Federal Statistical Policy and Standards (1980) Report on Exact and Statistical Matching Techniques. Statistical Policy Working Paper No. 5 Washington DC: U.S. Office of Management and Budget. (http://www.fcsm.gov/working-papers/wp5.html).

Office of the Inspector General, U.S. Office of Personnel Management (2011) *Stopping Improper Payments to Deceased Annuitants*. Washington DC: U.S. Office of Personnel Management, Office of the Inspector General, (September 14 2011). Available at http://www.opm.gov/oig/pdf/RP_Paper%209-14-11.pdf.

Overhage, J.M., Tierney, W. M., Zhou, X., & McDonald, C. J. (1997) A Randomized Trial of "Corollary Orders" to Prevent Errors of Omission. Journal of the American Medical Association, 4 (5), 364-375.

Peabody, J., Luck, J., Jain, S., Bertenhal, D., and Glassman, P. (2004) Assessing the Accuracy of Administrative Data in Health Information Systems. *Medical Care*, 42 (11), 1066 – 1072. (Abstract: get full article).

Pepper, J., Petrie, C., and Sullivan, S. (2010) Measurement Error in Criminal Justice Data. Chapter 18 of: A. R. Piquero and D. Weisburd (Eds) *Handbook of Quantitative Criminology*. New York: Springer, pages 353-374.

Petrila, John (2011) Law, Research, and Electronic Data Systems Used in Social Science Research. Paper presented at the Intelligence for Social Policy Conference, November 28-29 2011, Washington DC, University of Pennsylvania. Author: Petrila@usf.edu

Peugh, J. L. and Enders, C. K. (2004) Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*. 74 (4), 525-556.

Prewitt, Kenneth (2005) Racial Classification in America. *Daedalus*, Winter, 5-17.

Rescher, Nicholas (2007) *Error*. Pittsburg PA: University of Pittsburg Press (p 8).

Richman, Estelle (2011) Introductory Remarks. Conference on Intelligence for Social Policy, Washington, DC

Rossi, P., Lipsey, M.W., and Freeman, H. (2004) Evaluation: *A Systematic Approach*. (Seventh Edition). Thousand Oaks, California, London U.K, and New Delhi India: Sage Publication.

Taeuber, C. (2002) Developing community statistical systems with American community summary profiles and administrative records. *ASA Proceedings. Section on Survey Research Methods*. Retrieved May 25, 2010 from: jttp://www.amstat.org/sections/SRMS/Proceedings/Y2002/Files/JSM2002-000689.pdf.

Tippet, L. H. C. (1943) *Statistics*. London: Geoffrey Cumberlage Home University, Ltd. Oxford University Press.

U.S. Government Accountability Office (2009) Report 10-224T on the Recovery Act. http://www.gao.gov/new.items/d10224.pdf.

U.S. Government Accountability Office (2010) Report 10-437 on the Recovery Act. http://www.gao.gov/new.items/d10437.pdf.

U.S. Government Accountability Office (2010) Head Start: Undercover Testing Finds Fraud and Abuse at Selected Head Start Centers. Statement of Gregory D. Kutz. Washington D.C: U.S. GAO (May 18, 2010). (GAO – 10 – 733T).

U.S. Government Accountability Office (2009) *Health Information Technology*. Washington DC: USAO (Report 09-312T). Retrieved June 2, 2010, from http://www.gao.gov/new.itesd009593.pdf.

U.S. Government Accountability Office (2012) Federal Statistical System: Agencies Can Make Greater Use of Existing Data, but Continued Progress is Needed on Access and Quality Issues. Washington DC: USGAO (Report GAO-12-54)

Wang, S. J., Middleton, B., Prosser, L. A., Bardon, C. G., Spurr, C. D., Carchidi, P. J., Bates, D. W. (2003). A cost-benefit analysis of electronic medical records in primary care. *American Journal of Medicine*, 114, 397-403.

Wansbeek, T. and Meijer, E. (2000) *Measurement Error and Latent Variables in Econometrics*. New York: Elsevier.

Whitt, H.P. (2006). Where did the bodies go? The Social Construction of Suicide Data, New York City, 1976-1002. *Sociological Inquiry*, 76,166-187.

Wholey, J. S., Hatry, H. P. , and Newcomer, K. E. (Eds) (2010) *Handbook of Practical Program Evaluation*. Third Edition. New York: Jossey Bass/Wiley.

Wolff, N., & Helminiak, T. W. (1996). Nonsampling measurement error in administrative data: Implications for economic evaluations. *Health Economics*, 5(6), 501-512.

World Health Organization (2007). Estimation of Malaria Disease Burden in India: Report of an Informal Consultative Meeting New Delhi, India, 21–23 November 2007 http://www.searo.who.int/LinkFiles/Meeting_Reports_MAL-256.pdf

Zimring, Franklin (2011) Decline in Crime in New York City: 1990-2010. The Sherman lecture. Department of Criminology and the Jerry Lee Center for Criminology, University of Pennsylvania, Philadelphia PA (October 17)

Acknowledgements

I am pleased to have done this work with the support of the MacArthur Foundation. I am grateful to colleagues, John Fantuzzo and Dennis Culhane, for provoking the work, and to my students and some of my colleagues. The students include Elias Arellano who was industrious in his duties as my Trustee Chair Fellow, and to Yuefeng Xu, H. Hien, Jason Fischer, Rachel Anderson, Benjamin Brumley, and Xia Mu, who contributed nicely to my understanding during my courses. I am especially grateful to Mike Maltz for his tutoring me, briefly but informatively, in the matter of cop records. I am indebted also to anonymous reviewers of an earlier version of this paper. The mistakes and misunderstandings here are mine, not theirs.