Findings from the Integrated Data Workshops hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society

Alexandra Brown, University of Maryland[1]
Andrew Caporaso, University of Maryland
Katharine G. Abraham, University of Maryland
Frauke Kreuter, University of Maryland

Abstract:

Across agencies in the Federal Statistical System, there has been growing interest in integrating traditional survey and census data with auxiliary data (both structured and unstructured) in order to increase the quality and the timeliness of the data and statistics that are produced. To address the gap that exists in understanding the quality of such datasets (as compared to sample surveys) and improve communication between data producers and data users, the Federal Committee on Statistical Methodology Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources and the Washington Statistical Society) co-hosted three workshops that explored current practices with respect to reporting on the quality of integrated data. This report summarizes the three workshops and the key themes emerging from them.

[1] Corresponding author. Can be reached at: abrown53@terpmail.umd.edu.

Table of Contents

Introduction

Across agencies in the Federal Statistical System, there has been growing interest in integrating traditional survey and census data with auxiliary data from nontraditional sources in order to increase the quality and the timeliness of the agencies' statistical products. Data integration provides the opportunity to enhance existing statistical products with additional levels of temporal or cross sectional granularity. Data integration also may make possible entirely new statistical products that could not be created based on sample surveys alone.

The measurement of quality in traditional surveys and census data has been guided by the well-studied Total Survey Error (TSE) paradigm. This paradigm does not fully address many of the characteristics of non-survey data sources. To address the lack of a framework designed for thinking about the quality of datasets that incorporate information from multiple sources, and to improve communication between data producers and data users, the Federal Committee of Statistical Methodology (FCSM) Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources and the Washington Statistical Society (WSS) co-hosted three workshops that explored current practices with respect to reporting on the quality of integrated data. The first workshop focused on the quality of input data (December 1, 2017), the second on processing data (January 25, 2018) and the third on the quality of output data (February 26, 2018).

Two overarching questions were considered at each of the three workshops. These questions were:

1. *Fitness for use*: What quality features are important for data users to understand when either considering to use a particular integrated data source or evaluating the strengths and weaknesses of the resulting analysis?

2. *Communication with stakeholders*: What is the best way to communicate these predominant quality features to stakeholders, whether they be technical specialists, substantive data users, or the general public?

This appendix summarizes the presentations given at the three workshops. The recordings and presentation slides can be found at: http://washstat.org/presentations/.

Prior work on analysis with multiple data sources in the Federal Statistical System

The structure of the workshops was heavily influenced by The National Academies of Science, Engineering, and Medicine (NAS) report titled *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (available at: http://nap.edu/24893) (National Academies of Sciences, Engineering, and Medicine, 2017). This report explores the issues of moving beyond existing quality frameworks to fully address the quality and privacy considerations inherent with integrating multiple data sources.

During the first workshop, Robert Groves, of Georgetown University and Chair of the Panel that authored the NAS report, gave a presentation that focused on the findings from its Chapter 6. This chapter calls for a new, broader framework for assessing the quality of administrative and private sector data, such as organic or found data. Specifically, recommendation 6-1 calls for this broader framework to include additional dimensions that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency and interpretability. Recommendation 6-2 extends this to suggest that Federal statistical agencies outline the strengths and weaknesses of auxiliary data sources using the comprehensive quality framework so that these are transparent to data users. Agencies should focus on the tradeoffs between the different aspects of quality (such as trading timeliness for granularity) rather than focusing solely on accuracy in evaluating auxiliary sources.

Groves stressed the importance of an increased focus on the measurement of coverage error as the statistical agencies make increasing use of non-survey data. He also emphasized the need to understand the data generation process for these data in order to assess the construct-measurement gap, any linkage errors, and the overall quality of the resulting data. Ultimately, there needs to be effective communication with the general public around any sensitivities that may exist in such a way that they can trust the resulting data and statistical estimates.

Themes and Takeaways from the FCSM Working Group/WSS Workshops

The discussion during the workshops echoed many of the themes from the NAS report, including its emphasis on timeliness, privacy, and transparency. The first workshop was organized around identifying data quality standards and issues for non-survey data. The first session of this workshop focused on administrative data and the second on less-structured and unstructured datasets. An important takeaway from the sessions was that administrative data sources share many parallels with traditional survey data with respect to the assessment of data quality. This is less true of less structured and unstructured data sources.

Another important takeaway from the first workshop was the importance of clear and appropriate communication from data producers about the original reason for the collection of auxiliary data that are used for statistical purposes. Transparency around the strengths and weaknesses of the technology used to collect the auxiliary data was emphasized, with a focus on how changes in the data collection technology might affect the comparability of the data over time. Such transparency will give data users a better understanding of the auxiliary data sources, allowing them to better determine the error properties of a particular integrated dataset or survey statistic.

During the first workshop, Cohen illustrated the challenge of assessing administrative records' fitness for use by describing the numerous sources of health expenditure records, each reporting on different and sometimes non-overlapping types of healthcare data (e.g., charges, costs, and claims). Data processing, the topic of the second workshop, becomes increasingly complicated as more sources are introduced. Processing data sources into a unified format can

help to reduce issues related to measurement error, data coherence and comparability. Careful consideration of quality is needed for each individual data source before it can be deemed fit for a specific use. As an example, Muth discussed the differences in quality of scanner data for purchases of products with barcodes and products that are weighed (e.g., vegetables). Several of the presenters during the first workshop discussed one or another multi-faceted tool that has been used to assess quality trade-offs, such as the “The Data Quality Matrix” used by BLS to detail and weigh the quality implications of a particular source for various uses (see Murphy and Konny presentation).

Incorporating external data sources reduces the amount of control that the statistical agency has over data delivery, data quality and transparency. Agencies must trust that data providers will meet production expectations and also develop contingency plans in the event that data from a particular source cease to be available or cannot be delivered. Both Berning and Sheppard, from the U.S. Census Bureau, and Murphy and Konny, from the Bureau of Labor Statistics, pointed to the need to verify that a source will exist and remain consistent over time before it is relied upon for statistical production. Transparency in communication regarding data attributes can be compromised when data vendors are restricted in what they are able to share about a data source, whether for proprietary reasons, to minimize disclosure risk or to maintain the integrity and predictability of estimates.

Less- and unstructured data sources present additional processing challenges for researchers while offering the opportunity for increased timeliness and accuracy of data products. With regard to timeliness, Elkin discussed how quickly a combined database of health records and handwritten doctor notes can be searched to address doctors’ questions. With regard to accuracy, Johnson discussed being able to use combined data from multiple satellite images in order to better classify crops and point field staff to counties that may be most affected by weather. Mahapatra discussed accuracy in terms of analysts’ ability to view transportation routes holistically through the combination of multiple datasets that contain, for example, traffic, train and metro information. Rigobon discussed the strengths and weaknesses of using web-scraped data on prices in place of or in combination with survey-based price data. He determined that final data products combining multiple data types could have better coverage than products based on any single type of data.

The theme of the second workshop was data processing. Four sub-topics were covered: record linkage, statistical matching/data fusion, harmonization across data sources, and disclosure avoidance. Similarly to the first workshop, an overarching theme of the second workshop was the importance of transparent and clear communication with data users.

With regard to record linkage and statistical matching/data fusion, Steorts discussed the value of high quality “truth decks” (e.g., hand-matched subsamples) for determining the quality of a linked data product. Reiter focused on the importance of being transparent about modeling assumptions and matching techniques and the biases these might create in linked data products. Reiter stated that Bayesian multivariate models are a promising theoretical framework

for combining datasets in the non-record-linkage realm, but that these models do not yet incorporate an understanding of quality profiles for different data sources.

One implication coming out of the discussion during this part of the workshop is that survey designers should plan for data harmonization as a fundamental first step in integrating multiple data sources; in many instances, entity resolution and statistical matching would not be possible without effective data harmonization. Jang recommended that harmonization be treated as its own separate process with appropriate resource allocation throughout the research process.

The work that Holan described relates to the development of estimation techniques for use when sources change in their spatial and temporal properties. This line of research is relevant to facilitating harmonization when data sources do not align perfectly. Reist's exploration of using survey data to assess coverage and measurement error properties in administrative record sources is another useful line of research.

A theme discussed throughout the second workshop was the importance of data privacy and the ethical treatment of data. Sweeney reported that, under current data delivery standards, disclosure risk remains unacceptably high. One disclosure avoidance technique that offers an alternative to cell suppression is to add random noise to overcome the consequences of the database reconstruction theorem. With this approach, it is possible to be transparent both about the methods used and about the error properties of the resulting estimates. Sweeney cited the Commission on Evidence-Based Policymaking Final Report (2017) for guidance on disclosure avoidance.

The third workshop focused on the quality of output data. Data products may have users of varying sophistication and the level of detail provided for one type of user may not be appropriate or necessary for another user. Organizers discussed different transparency reporting levels that could be used to determine the amount of documentation provided alongside a final data product. The levels of transparency discussed can be summarized as follows:

- High Transparency, for academics, agency specialists, subject-matter experts
- Moderate Transparency, for policy makers, professional journalists, students
- Low Transparency, for the general public

Langton's presentation and Eltinge's accompanying discussion illustrated the similarity between methods for identifying and addressing breaks in time series for survey data and integrated data.

Several models for reporting on the quality of integrated data that have been developed outside of the U.S. Federal statistical system were presented during the workshop. Biemer presented a way to expand the TSE model, developed originally for the evaluation of survey data, to incorporate datasets and hybrid estimates based on integrated data. Tying this work back to transparency, Biemer stressed that data users need to be able to follow the logic and assumptions made in developing a data product in order to determine its fitness for use. Czajka

reviewed quality assessment models developed elsewhere in the world, contrasting these models with the quality profiles that at one point were thought by some in the U.S. statistical system to be the way forward. This set up a discussion about whether quality profiles should be revived for reporting on the quality of integrated data or whether it would be better to adapt a framework utilized by another country to fit U.S. needs.

Kreuter summarized and concluded the workshops with a call to rethink existing research paradigms in the world of integrated data. She pointed out the need to assess quality at the estimate level and noted the challenge this presents, given that each individual estimate may require its own tailored quality assessment. She also suggested that interdisciplinary research collaborations should be promoted, noting that no single discipline is likely to possess all of the expertise required to develop and work with integrated data products.

Overall, through the course of the three workshops, the strongest and most often repeated recommendation coming from the speakers and discussants was for data producers *to be transparent*. Transparency is needed at each step of data product development, from the original motivation for collecting the input data, to the steps that were taken to harmonize multiple data sets, to the matching procedures and model assumptions chosen, all the way through to the evaluation techniques the researchers used. Where possible, it is also important to be transparent about how entity privacy will be maintained in the datasets. Where it is not possible to be transparent about disclosure avoidance techniques, the preservation of privacy still should be a priority for data producers. Data processing decisions for every step of the collection and analysis process should be documented so that stakeholders can judge whether a given product is appropriate for their research.

When thinking about transparency, it is important to think about the target audience for the information. During the workshop discussions, there were multiple references to levels of transparency and tiered access for data users depending on their level of sophistication. These suggestions were made both to allow users to find the information they needed in the format that worked best for them, but also to create an environment where disclosure risk could be minimized while maintaining effective communication with data users.

Transparency is particularly important in light of another theme heard throughout the workshops, namely, that fitness for use is apt to be very project specific. For example, if there is a particular need for information on a short timeline, then it may be acceptable to trade off accuracy for more timely data delivery. This same trade-off might not be acceptable, however, for an end user who is focused on the most accurate data product.

Merging and processing multiple data sources requires greater coordination and additional resources relative to conventional survey data processing. Input data are not likely to be in the same format as the survey metrics of interest, which means they must be transformed before they can be used. Processing includes the careful work of data harmonization and statistical matching, both areas in which there is considerable scope for future research. On the backend,

once data have been processed, it is important to share with the end users what was done to prepare the datasets to be integrated while also preserving the identity of the underlying entities.

Lastly, in terms of moving forward with integrated data work, a unified research plan is needed along with more communication between the federal statistical system and stakeholders as to the major priorities for facilitating the use of integrated data products.

Report Outline

This report will continue with a summary of each of the presentations at the three workshops. Additionally, there was a follow-up session at the 2018 FCSM conference that asked the presenters and audience members to reflect on the research priorities and questions that came out of the workshops. A summary of this session follows the workshop summaries.

Workshop 1: Quality of Input Data

Workshop 1 was developed to identify quality issues related to non-traditional data themselves before they are blended with survey data. The first part of this workshop focused on administrative data sources that typically are structured in much the same way as survey data. The second part of the workshop focused on semi-structured data such as those collected through automated sensor systems, and unstructured data such as those captured through images or raw text. Table 1 below illustrates how different types of survey and non-survey data fit into these three categories—structured, semi-structured and unstructured. Each part of the workshop included four speakers from the federal statistical community who shared their experiences with assessing quality when leveraging auxiliary data sources.

Table 1: Examples of Structured, Semi-Structured, and Unstructured Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

Source: Groves et al., *Innovations in Federal Statistics* (2017)

Session One: Structured and Administrative Data Sources

Speakers:

1. Steven B. Cohen, RTI. *The Utility and Limitations in Administrative Data for Medical Care Expenditure Analysis*
2. Michael Berning and David Sheppard. U.S. Census Bureau, *Quality of Administrative Records as Source Data*
3. Bonnie Murphy and Crystal Konny, Bureau of Labor Statistics. *Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) and Consumer Price Index (CPI) development*
4. Mary Muth, RTI. *Assessment of Commercial Store and Household Scanner Data: Methods, Content, and Cautions*

Steven B. Cohen, RTI – "The Utility and Limitations in Administrative Data for Medical Care Expenditure Analysis"

Steven Cohen outlined how administrative record data on medical expenditures can be incorporated into survey data analysis. Health expenditure data are used for a number of reasons, including to inform health policy, identify health disparities, measure the burden of chronic conditions and also to estimate the proportion of gross domestic product (GDP) attributed to health expenditures. Integrating administrative record sources on expenditures with survey data presents an opportunity to facilitate deeper exploration of rare conditions and to do so with reduced burden on the respondent. Using records also provides a source for benchmarking survey estimates and allows for research into the quality of survey estimates. In the context of a longitudinal survey design, access to health records over time could help facilitate 'signal detection' (impending health issues or hidden patterns). Integrating large administrative sources with representative survey data can facilitate 'hybrid' designs that leverage the representativeness of the survey data with the breadth and depth of administrative sources to produce estimates that are nationally representative at a lower cost than attempting to get those estimates exclusively from survey data.

There are a number of public and private data sources for medical expenditure information. The sources vary in who they cover, what data they collect, and how accessible or expensive they are. The Healthcare Cost and Utilization Project (HCUP), managed by the Agency for Healthcare Research and Quality (AHRQ), is an effort to provide standardized and representative expenditure data at the national level; it is currently the largest collection of longitudinal hospital discharge data in the United States, representing 98% of all discharges. Similar efforts to the HCUP have been made at the state level to produce All-Payer Claims Database Systems (APCDs). The Medicaid Analytic eXtract (MAX) includes person-level data on eligibility, utilization and payments for the Medicaid population. An example of a private source is the MarketScan Data Warehouse, which includes patient and claims data from a

convenience sample of large US employers. Another private source, Optum Labs Data Warehouse (OLDW), houses claims and clinical data from over 100 million individuals. Medical expenditure data can also be made available through the medical billing systems of healthcare companies. However these are typically not standardized from one company to the next. Cohen points out that, as electronic health records (EHRs) become more standardized and computerized, they will become a rich source of health and health expenditure data that could supplement or replace survey efforts.

A number of factors are considered when determining the quality and fitness for use of a particular data source, including the purpose of the data generation, degree of documentation, and compatibility with the current research objective. Regarding compatibility, researchers must consider whether the source is reporting on costs, charges or payments and whether these types of data can be leveraged for their analysis. Cohen stresses that understanding who is represented in the source is critical for determining fitness for use. In addition to assessing quality and fitness for use, consideration must be given to whether incorporating additional data sources puts study participants at risk of disclosure.

Cohen describes the Data Integration Model, which is a process of unifying multiple related data sources to enhance their analytic utility. The resulting integrated data can be used for a number of purposes, including to inform sample design, nonresponse and imputation, and to reduce measurement error. The backbone of an integrated data model could be a survey frame or a rich administrative data source that would be used as the basis for linking to other primary and secondary data sources. Cohen emphasizes the need to formalize sources into a unified structure (if, for example, one reports on charges and another reports on payments) within the model in order to facilitate use for estimation. Decisions must also be made as to which source will be used when multiple sources are available, especially if reports from the sources do not align.

Cohen points to two examples of administrative records being incorporated into current health expenditure research. The Medical Current Beneficiary Survey (MCBS), conducted by the Centers for Medicare and Medicaid Services (CMS) uses Medicare enrollment files as the basis for its sampling frame. Using this source facilitates assessment and improvement of the accuracy of survey reports, The Medical Expenditure Panel Survey (MEPS) is an annual survey of 14,000 households that serves as the ‘quintessential’ design for merging survey and administrative record sources. MEPS includes a Medical Provider Component that seeks to collect data on the respondent directly from their provider in order to improve the accuracy and completeness of data on expenditures and utilization.

Michael Berning and David Sheppard, U.S. Census Bureau - “Quality of Administrative Records as Source Data”

Michael Berning and David Sheppard describe the Census Bureau’s experience and ongoing efforts to incorporate and assess the quality of administrative records. The newly developed

Data Acquisition and Curation (DAC) area within the Economic Directorate (formally a part of the Center for Administrative Records Research and Applications (CARRA) within the Methodology Directorate) was recently established to help manage and assess the quality of the growing number of outside data sources used by the Census Bureau. DAC manages over 150 interagency agreements to share administrative records and outside survey data sources. The Census Bureau has the authority to request data from other institutions but these institutions are not necessarily required to provide the data. Therefore the DAC will typically accept data in whatever format it is provided in order to maintain a positive relationship with the provider.

The authors note that with the increased number of administrative records available to the Census Bureau, the need to distinguish between high and low quality sources ("signal and noise") has become more important. The increase in sources has also led to an increase in potential users of such data; within the Census Bureau, the data are used to help with operations related to the Decennial Census and other surveys as well as for research purposes.

The DAC evaluates each source using a data quality framework that assesses several qualitative and quantitative factors relevant to using the data. The qualitative factors stem from Iwig, Berning, Marck and Prell (2013): relevance, accessibility, coherence, interpretability, accuracy and institutional environment. The authors emphasized the replicability of administrative data sources as a component of their relevance: It is important to determine whether a given source will be available and consistent over the timeframe that it will be used. Accessibility can be impacted by laws governing the use of data or the cost of acquiring the data. Coherence can be an issue when getting data at the state level, as each state may have different methods of keying data. The quantitative factors include metrics such as the degree of missingness and frequency of invalid values. Cost effectiveness is also evaluated before incorporating a particular source.

Regarding past experience in assessing data quality, the Census Bureau has maintained a contact frame stemming from Decennial census data that can be used to evaluate outside data sources before they are used for analysis. In requests for proposals (RFPs) to provide data, the Census includes requests for sample data that they can compare to the contact frame in order to assess the quality of a particular vendor's data. Administrative record sources of teacher rosters have been used by the Census Bureau as a preliminary source for developing school-level rosters; sampled schools are asked to verify a prefilled roster instead of producing it from scratch, thereby reducing their burden while at the same time evaluating the quality of the source. Administrative records have also been directly compared to American Community Survey data to assess their quality.

The DAC's Data Quality Branch is currently working to establish a documented and repeatable process for acquiring and ingesting new data sources while ensuring data quality. The authors stressed the importance of communication with providers in order to facilitate cooperation and timely data delivery that meet users' expectations and reduce duplication of effort. It's common for the first data delivery from a particular source to fail to meet expectations and upfront communication can help minimize this risk. Communication with data users regarding quality is

also a priority of the Census Bureau and DAC. The DAC aims to share more metadata on data quality and expand their data quality review process by comparing source data to benchmark statistics.

Bonnie Murphy and Crystal Konny, Bureau of Labor Statistics – "Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) and Consumer Price Index (CPI) development"

Bonnie Murphy and Crystal Konny summarize how the Producer Price Index (PPI) and Consumer Price Index (CPI) use administrative ("alternative") records to facilitate data collection and analysis. Both the CPI and PPI have used administrative records for decades to improve and augment price statistics. They utilize publically available data (for example, Department of Transportation data), purchased data (JD Power), and company provided data (either provided directly, through web scraping, or through an Application Program Interface (API) when available and permitted).

Administrative data are used for a wide variety of purposes including to create sampling frames, or for supplementing, validating or imputing survey data. The potential benefits of using administrative records include reduced costs and respondent burden and increased sample size, precision, and timeliness. With respect to creating sampling frames, Murphy explains that certain data on expenditure categories are more efficiently obtained from administrative data sources – for example, JD Power data are more efficient source of data for used car sales than the CPI's standard source. The PPI uses administrative records for data on expenditures in the service industry where employment is not a good measure of the size of the industry.

The BLS carefully examines a number of data quality metrics when determining if a source is fit for a particular use; the metrics are similar for any administrative data source they consider. Sources are evaluated over a long period of time to ensure they meet coverage, accuracy and reliability requirements.

Accuracy is both the most important and most challenging quality dimension to measure when determining the fitness for use of a given source. Transparency matters in several ways: (1) transparency of the data (does it have unknown estimation techniques or biases?), (2) transparency of PPI and CPI (will divulging sources make the PPI and CPI predictable?) and (3) transparency implications for maintaining respondent confidentiality. Using outside sources takes quality control out of the hands of the BLS, which limits transparency and the ability to measure error.

Coverage is an important accuracy dimension to consider with respect to time, geography, outlet/establishment, and item. Item-level purchase data are the most important for data quality; sources can vary in whether they report producer or consumer prices and whether prices include discounts or sales. Sources must also provide sufficient information about products in

order for the BLS to monitor price change. For example, a company may change the size of a product (for example, reducing the size of a box of brownie mix from 19.8 to 18.3 ounce), but this type of detail may be difficult to capture from administrative data sources.

Incorporating administrative data from an outside source may jeopardize the timeliness of data delivery, as data procurement at the BLS can be a slow process due to internal rules and regulation. Relying on outside sources also takes control of data delivery out of the hands of the BLS, which limits internal control of data quality. Additionally, BLS contracts can only span 5 years, after which time data from a contracted source may change in quality or price.

An outside supplier of data must be reliable—the BLS must be able to count on the supplier to exist moving forward, provide cleaned and validated data and remain secure. Regarding security, the BLS needs to be sure that an outside supplier will not attempt to manipulate the CPI and PPI if they know that their data are being used for those programs. Using outside data sources may impact the accessibility of the indices or at least the data underlying the indices, as some sources may restrict who is allowed to view and use their micro data.

Konny stresses the importance of considering costs of incorporating a data source when evaluating its quality. Costs must be considered with respect to purchasing the source itself as well as adjusting BLS infrastructure and resources to accommodate use of the source. Resources should be allocated for contingency plans in the event that data from an outside source become unavailable.

Konny describe the Alternative Data Matrix, which is designed to weigh impacts of using an administrative data source based on a number of data quality dimensions. The matrix offers a concise way to assess fitness for use on a number of dimensions and communicate quality considerations. With respect to communication about administrative record quality, the BLS must weigh being transparent about methodology and protecting the confidentiality of data suppliers.

Mary Muth, RTI – “Assessment of Commercial Store and Household Scanner Data: Methods, Content, and Cautions”

Mary Muth summarizes considerations related to the utility and quality of scanner data for research on food expenditures. Scanner data collection involves scanning barcodes on products in stores to track information on who purchases them and in what quantity. Scanner technology allows for collecting rich, high-volume expenditure information in a low-burden way relative to conventional self-report methods. Scanner data can be used to summarize expenditure information at the household, product, store, and geographic level.

The Economic Research Service (ERS) has incorporated scanner data to construct prices for the Economic Research Service (ERS) Quarterly Food at Home Price Database. Scanner data are also used in calculating the cost of the WIC food package and the cost of the Thrifty Food

Plan, which is the basis for the SNAP allotment formula. Muth pointed out two reports from the ERS and US Department of Agriculture (USDA) that address the quality of scanner data and compare it to government survey data (see citations below).

Private companies such as IRI and Nielsen collect and provide scanner data on expenditures, primarily for commercial purposes. As a consequence, the utility and transparency of the data for research purposes are somewhat limited. For example, the providers offer limited documentation on sampling, data collection and weighting methods. Muth notes that the cost of purchasing a license to use the data is considerable. The data are collected in two ways: (1) store-based scanner data collected directly from retailers who provide aggregated sales statistics and (2) household-based scanner data collected from panelists of the National Consumer Panel (NPC) who report on their expenditures throughout the year. Muth details a number of quality considerations with respect to both methods of scanner data (store-based and household-based) provided by IRI.

Store-based scanner data. IRI's store-based data represent a sample of over 59,000 stores that have agreed to provide sales data from all of their locations. This sample is estimated to represent 41% of stores and 55% of food sales based on 2012 Economic Census data; these rates vary somewhat by the type of store (convenience, grocery, etc.). The data are not designed to represent smaller, independent stores. IRI aggregates sales for individual products and reports weekly sales.

Stores vary in what they report, for example some stores do not provide data on their store-brand products. Some stores aggregate sales from all of their locations while others provide data for individual scores or regions. Stores vary in how they report prices with respect to coupons and loyalty card discounts.

Sales of random-weight products (for example, produce or deli meat) are more difficult to capture with the same level of detail as products with Universal Product Codes (UPC). Stores vary in how they sell and report purchases of random-weight products, which limits the utility of scanner data for analyzing them.

There are restrictions in what types of information derived from the scanner data can be reported (for example, brand names are not allowed to be reported). To minimize disclosure risk, IRI will not provide projection weights for individual store data.

Household-based scanner data. Household-based scanner data have a different set of quality considerations. NPC respondents are recruited online and then randomly selected to participate based on demographic quotas. Selected households record purchases using an in-home scanner or mobile app.

Approximately 126,000 households participated in the NPC in 2012. Only households who meet minimum participation and purchasing requirements are assigned projection weights. In 2012 about half of NPC households met the criteria for inclusion. Households that failed to meet

qualifications were more likely to be younger, lower income, black and Hispanic, and with children. Muth notes that qualifying households may differ from the general public given their willingness to report all of their purchases; experience in reporting purchases also leads survey participants to become more aware consumers, which may affect their purchasing behavior.

Participants are asked to report the store where they purchase each product and whether the product was discounted in any way. IRI attempts to assign a price for the transaction based on average prices for the chain or other stores in the market area; they will only ask the respondent to report the price if such data are not available. For random-weight products the process is more burdensome and less detailed; the respondent must select the type of product from a list and enter the total price paid without reporting the quantity.

The household data are weighted using Iterative Proportional Fitting (raking) to demographics (not expenditure or shipment data). In comparing 2012 NPC data with the Consumer Expenditure Survey and FoodAPS (two large survey sources of expenditure data), Muth reports that agreement between the IRI data and these survey sources range from less than 50% to over 80% for different types of products. Products that are typically random-weight tend to be underrepresented in the NPC data. IRI has methods for adjusting for the undercount of random-weight products but the details of the methods are not released for proprietary reasons.

Session Two: Less structured Data Sources

Speakers:

1. Dr. Peter Elkin, SUNY Buffalo. *The Improvement in Sensitivity and Often Specificity when Adding Unstructured to Structured Data*
2. David Johnson, USDA, National Agricultural Statistical Service. Data *Quality of Satellite Imagery for Studying Complex Systems*
3. Subrat Mahapatra, Maryland Department of Transportation. *Sensing Data Quality in Sensor-Based Data*
4. Roberto Rigobon, Massachusetts Institute of Technology. *Web-scraped Data, Consideration of Quality Issues for Federal Statistics*

Dr. Peter Elkin, SUNY Buffalo - "The Improvement in Sensitivity and Often Specificity when Adding Unstructured to Structured Data"

The first presentation was given by Dr. Peter Elkin, MD, MACP, FACMI, FNYAM, SUNY Buffalo. Dr. Elkin is the Director of Informatics Core of the University of Buffalo Clinical and Translational Science Institute and a Professor and Chair in the Department of Biomedical Informatics.

The data that Dr. Elkin uses in his work are electronic health records. These records have a long history, and, over time, standards for how they should be organized have developed. Dr. Elkin indicated, however, that these standards are fully implemented only 18% of the time. In his view, we need to do better.

To improve the consistency of electronic medical records, Dr. Elkin and his team have combined structured and unstructured data to develop an ontology-based data system. Ontology is defined simply as a systemized way of naming things that also has an infrastructure that is computable. Within this system, descriptive logic can be used to create formal definitions that are computable and give researchers hierarchies based on the terms in the definitions. This has the benefit that, when researchers or practitioners come to the same problem from different perspectives, they all will have consistent data.

The unstructured data were prepared for the database by parsing them into structures through natural language processing. When combining the unstructured data with the structured data, the developers recognized the importance of looking closely at data quality and accounting for missing and duplicate data as well as examining time series of events to identify erroneous or misfiled data. Through their data quality explorations, the developers determined that validation data were critical to developing an accurate and functional database.

A number of successes resting on the combined datasets have been reported. First, doctors now receive quality-of-care report cards with much of the reported information coming from

unstructured data that previously were unused. There also has been an improvement in the quality of care as measured by metrics of practice outcomes, including an improvement in the number of internal referrals seen (from 54% to 82%), an indication that doctors in different areas of specialty are now working more closely together.

To sum up this on-going project, Dr. Elkin pointed to his research entitled “AI and Natural Language Processing (NLP) to Enhance Structured Data’s Ability to Identify Nonvalvular Atrial Fibrillation Patients and their Stroke and Bleeding Risk” and “Strokes Prevented: Biosurveillance of NVAF Patient Cohorts CHA_2DS_2-VASc and HAS-BLED Scores Using Natural Language Processing and SNOMED CT”. In this research it was determined that structured and natural language processing combined were statistically better than just structured data alone. Dr. Elkin’s argument is that if structured and unstructured data can be combined effectively, more people can be treated, many deaths can be prevented, and a lot of money can be saved.

David Johnson, USDA National Agricultural Statistics Services - “Data Quality of Satellite Imagery for Studying Complex Systems”

David Johnson, of the National Agricultural Statistics Services, was the second presenter of the session. His presentation showed how satellite remote sensing imagery can be used separately or can be combined with additional data sources to add information to regular crop production and area reports. These unstructured data sources, consisting of spatial images, can improve both the quality and the timeliness of the data in the reports. There are two predominant uses of these data: land-cover mapping and yield estimation.

Land cover mapping looks at what crops are growing where. The Landstat program has been using satellite images for the last 45 years to gather these data, and the data are made publicly accessible without charge at https://glovis.usgs.gov/ . In order to process the satellite images, the images (the input data) are gathered, a sample is selected, and a decision tree is utilized to extrapolate out to the areas that aren’t covered in the image. The area is then classified with the different crops expected to be growing there, and finally this information becomes output as an ArcGIS file. The accuracy of these classifications is tested through comparisons with administrative data, developing the classifications with a portion of the data and then seeing how well the left-out data are classified. This is done through comparing the classified data to enumeration data (regression analysis); and through comparisons of different satellite images. On the yield side, daily images are captured through Modus. Modus has a 20-year history and the images are accessible at www.worldview.earthdata.nasa.gov.

The Normalized Difference of Vegetation Index (NDVI) is a commonly used metric for distilling the massive amount of data collected through spatial images into a simpler and more digestible format. This metric is a proxy for how much biomass is in the landscape. Through the use of current and historical images, the NDVI can be used in modeling relationships among weather, location, and plant growth. Such models may make it possible to quickly predict the effects of a real time situation; for example, looking at past drought years can help to predict the effects of

current drought, while the real-time large datasets are being built and analyzed. Further, these models are built for the national and state level and can be updated in a short amount of time. A user-friendly tool that incorporates current NDVI metrics can be found at, https://glam1.gsfc.nasa.gov/.

The overall strengths of remote sensing are that when looking at land (not demographics or other aspects of the economy) there is good aerial (and global) coverage; the data are available at a good temporal rate (for example, daily or every 16 days); much of the data is free or is becoming free; there is little data latency (data that are captured today are available either the same day or the next day); there is fine spatial detail; the simple statistical models seem to be as good as complicated ones; and, with technological advances, better sensors are on the way and computing costs are falling.

There are weaknesses of remote sensing as well. It is still computationally intensive and requires an integrated skill set (e.g., a specialist needs to know how to code as well as something about geography and agronomy). The datasets and images are occasionally recalibrated, making it necessary to rebuild the models. The measurement of uncertainty within the data is not well established and some measures, such as temperatures, are estimates instead of precise readings, which creates noise in the data. Lastly, there is not a very long history (in relation to the history of USDA) of such images being used.

Subrat Mahapatra, Maryland Department of Transportation - “Sensing Data Quality in Sensor-Based Data”

Subrat Mahapatra, of the Maryland Department of Transportation, was the third presenter. His presentation focused on how both structured and unstructured Maryland transportation data are being used to gain an understanding of 21st century transportation needs. As Mahapatra noted, transportation needs and behaviors are changing and the technology of transportation is changing as well.

Using both types of data, along with new technologies to collect the data, allows for a holistic view of transportation that can be used to improve the coordination between real-time operations, planning, and maintenance decisions (i.e., moving away from traditional silos). The Maryland Department of Transportation is developing these next-generation datasets by shifting away from a focus on how a single intersection may be negatively affecting traffic to a focus on how that intersection affects a trip from point A to point B, in relation to the other intersections (or aspects) of the trip. The shift places an emphasis on accessibility - what can an individual get to within a half hour, combining multiple modes? This view helps to identify how consumers are impacted by the current system and improves the reliability of trip time estimates. The new data collection and dissemination technologies allow for better performance management and better communication with the customers so they are part of the solution. For the Maryland Department of Transportation, understanding the customer experience, how it is changing, and how the Department of Transportation is making a difference (or not) is a key priority.

The majority of the data that are managed by the Department of Transportation are structured, which means they are machine readable, can be stored easily in relatable databases, are standardized in some format so crosstabs and queries can be run, and follow some agreed-upon rules. Data are gathered from the Census and various surveys for planning and gathered from signals and sensors for operations. Sensor data are collected 24/7 365 days a year. Traditional sensors are static and capture a point-in-time image. Probe technology focuses on the speed and volume of vehicles. This mix of data is organized into input and output data streams so that more informed decisions can be made. For example, a model that has been proven to be accurate and to assist in improving the reliability of travel time estimates combines probe data with other data to create data streams that produce the highway signs reading “XX miles in XX minutes” posted along many major roads. Models using probe data have proven to be more accurate than models using sensor data since speed is not captured with sensors. In order to develop these data streams, a data governance principle is applied that includes reviewing the data for accuracy, relevance, timeliness, coherence, and comparability.

The unstructured data that the Maryland Department of Transportation uses include qualitative data, but it is still unclear how to translate social media and how social media can inform planning decisions. What social media do provide is an opportunity to look at the customer side of things. For example, they can be used to address questions such as, what is the population saying about the current strategy? While in the early stages, there are efforts to turn some of the unstructured data into structured data. Overall, the future will involve making use of both structured and unstructured data for multiple purposes.

Roberto Rigobon, Massachusetts Institute of Technology - “Web-scraped Data, Consideration of Quality Issues for Federal Statistics”

Roberto Rigobon, a faculty member at the Massachusetts Institute of Technology, founder of the Billion Prices Project and co-founder of PriceStats, was the last speaker of the session. Rigobon’s presentation discussed how web-scraped data can be used in the context of national statistics offices as well as the advantages and disadvantages of using such data. One of the reasons he generates and uses web-scraped data is to assist in his current work of studying international pricing practices and producing alternative measures of inflation.

To begin his presentation, Rigobon briefly described web scraping as the process of collecting data from the documents, images, and descriptions that websites provide. When data users are considering working with web-scraped data, they should be aware of the advantages and disadvantages of using these input data in order to determine if they are appropriate for the intended purpose, even before integrating them with other data sources. The advantages of web scraping are that it is a non-intrusive data collection method; that programs can be written relatively easily to read in exactly the information the researcher would like to collect; and that, because the data collection process is automated, it is quick and inexpensive.

There are also disadvantages to using web-scraped data. The first is that the data might not be representative. Users should try to understand what or who the data are representing. There might also be sample selection problems. Third, there is a concern over the reliability and consistency of the data that are available on the web. For example, if a national statistical office is collecting data from a company that has been around for 10 years, how do they know they can continue collecting these data for another 20 years? Further, websites themselves tend to change their structure (in addition to their content) regularly. Every time such a change occurs, the mapping that was built for the old site is no longer relevant and needs to be rebuilt.

The fourth disadvantage is that users need to pay close attention to how the data were created and whether or not they were treated by someone else. Treatment could include data aggregations that do not clearly indicate how many people are in each cell, who is represented in the data, and when the data were collected. The documentation for the data should also clearly indicate how it was collected so that users can better judge its usefulness for their work. For users, collecting pure data is ideal but it can be difficult to tell which data are “pure”. For example, it may be unclear whether a glowing review was written by an individual hired by the company or written simply because a user had a positive experience. If issues are identified, the data can be corrected, but if and how this correction is done needs to be very clear so the final data users can judge the correction criteria and determine the quality of the data for their own use. Fifth, in terms of e-commerce data, it can be hard to determine how transparent a company is being. Some companies will clearly indicate when an item is out of stock or on sale, but others do not, meaning that the prices on their websites may not be true transaction prices. Sixth, web-scrapers need to be cautious about protecting the privacy of the company whose website was scraped. Lastly, there is a concern that the consistency and reliability of the data may change over time. These changes occur when consumer behavior and consumers’ reasons for using certain websites change.

Once a website has been scraped and the data are available to the researcher, what can they be used for? Rigobon approached this question from the perspective of statistical offices that are trying to compute economic statistics or understand the economy, as opposed to the perspective of predicting GDP or trying to understand the behavior of competitors. The first, and possibly best use, of web-scraped data is to complement the surveys and statistics that already exist. The new data should not necessarily be a primary source, but instead may be used to fill the holes created by, for example, nonresponse. It is important to understand the representativeness of the web-scraped data with respect to the items within a given category that are actually purchased by consumers and the prices actually paid for those items. Even if the web-scraped data are not fully representative, they still may help to improve estimates for a specific subdomain.

Web-scraped data also can be used to make estimates and to forecast, for example, prices, without combining them with surveys and statistics that already exist. Web-scraped price data provide an illustration of when the disadvantages of the data matter more and less for the end product. To create single point-in-time estimates, reliability matters less since the researcher is looking to create an estimate that won’t necessarily be replicated in the future, diminishing the

fear of a change to the data collection process. For forecasting, however, reliability is very important, since the researcher needs the data collection method to remain the same over time so that the estimates can be replicated in the future. Lastly, web-scraped data can be used for measurement. Similar to Rigobon's own work, this could include measurements of exchange rates and inflation based on web-scraped data alone. For those sectors that aren't represented online, however, there will be missing information unless the web-scraped data are combined with data from other already existing sources. Rigobon argues that measurement may impose higher requirements than other uses and recommends against using web-scraped data to replace current statistics.

Workshop 2: Processing Data

Workshop 2 focused on identifying quality issues related to how input data are processed into an integrated data set and how the integrated data are structured. This workshop was divided into four sessions with each session focused on a specific component of data processing: 1) record linkage, 2) harmonization across data sources, 3) statistical matching/data fusion, and 4) disclosure avoidance. These four components were chosen because they were high priorities and were not going to be covered by the other workshops.

Session One: Record Linkage

Chair: Joe Schafer, U.S. Census Bureau
Speaker: Rebecca Steorts, Duke University
Discussant: William Winkler, U.S. Census Bureau

Rebecca Steorts, Duke University and U.S. Census Bureau – "Entity Resolution: Measuring and Reporting Quality"

Rebecca Steorts began her presentation by defining entity resolution as the practice of joining multiple data sets by removing duplicate entities, often in the absence of a unique identifier.

The big questions in entity resolution surround discerning whether the entity matched across two datasets is the same, how these matches can be made in a quick and automated way, and what metrics should be used to evaluate the quality of the matches. As an example of why entity resolution is so difficult to accomplish, Steorts presented a model in which there are a total of M records in D data sets. To conduct the entity resolution between these datasets, the researcher is looking for a model that (1) is much less than quadratic, and that also (2) is reliable and accurate, fits the data well, and accounts for the uncertainty in the model. Steorts stated that these desired model components tend to work against each other, so it is very difficult to have both at the same time.

One approach to entity resolution is de-duplication. Duplication can occur within a single dataset or across multiple datasets. In the latter case, de-duplication is conducted in two stages—first, all of the datasets are combined into a single dataset, and then the records in the single dataset are de-duplicated. Because the record linkage uncertainty cannot be propagated exactly, however, information can be lost through this process. Common examples of de-duplication methods include logistic regressions, random forests, support vector machines, Bayesian adaptive regression trees (BARTS), and locality sensitive hashing.

The second common approach to entity resolution is record linkage. With this approach, the researcher reviews the record linkage uncertainty of an entire graphical structure. Blocking is commonly used to simplify the record linkage task. Record linkage problems require a quadratic

number of comparisons. Reflecting on the first model goal listed above, if a partitioning of the data into non-overlapping blocks can be accomplished, then the computational requirements will be much less and the model will be less than quadratic. The most common methods used for blocking include deterministic and probabilistic approaches.

Deterministic blocking is based on features available in the database, such as gender. Whatever feature is used to form the blocks is not incorporated into the model, which results in computational savings and a quicker speed. Once a deterministic blocking feature is fixed, however, error cannot be propagated from it. Probabilistic blocking is based on locality sensitive hashing. An example of this is choosing a block through the use of a probability mechanism (e.g., nearest neighbor), accounting for the computational costs of the chosen mechanism.

With both approaches to entity resolution—de-duplication and record linkage—it is common to match on a unique identifier (e.g., Social Security Number) if it exists. Problems occur when the unique identifier is missing or has noise in it. Exact matching is also commonly used. With exact matching, the first record is compared to a second record to determine if all the features of the records match. If they match exactly, they are considered a match. The requirement for how closely these records need to match can depend on the evaluation method chosen. The non-uniformity of the various evaluation methods and the quality differences among them have pushed researchers to want a more systematic method like likelihood ratios or hypothesis tests with clarity around the threshold for acceptance.

The Fellegi and Sunter (1969) method is an example of matching that rests on a simple likelihood ratio test. This method compares two databases with an all-to-all comparison of records and uses a Neyman-Pearson hypothesis test with a predetermined threshold, t. For this test, if the relative likelihood of observing the pair formed by record i (from the first database) and record j (from the second database) when the pair is a true match as compared to when the pair is not a true match lies above t, then they are deemed to match. It is not clear, however, what this threshold should be. Additionally, this method is computationally intractable and transitivity is not preserved in applications with multiple data sets (in other words, if 1 matches 2, and 2 matches 3, then 1 does not necessarily match 3).

Other evaluation metrics for the performance of a particular record linkage include recall and precision, reduction ratio, estimated sample size, standard error of estimated sample size, computational run time (and complexity), robustness of tuning parameters, and a review to determine if supervised methods are overfitting the data. Descriptions of these metrics are below:

Evaluation Metric	**Description and Notes on Metric**
Recall (see abbreviations table below)	Recall = TP/(TP+FN)=1-FNR (estimates closer to 1, with FNR closer to zero, are preferred)
Precision (see abbreviations table below)	Precision=TP/(TP+FP)=1-FPR (estimates closer to 1, with FPR closer to zero, are preferred)
Reduction ratio	Measures the relative reduction of the comparison space from the de-duplication or hashing technique (see Steorts, Venture, Sadinle, and Fienberg, 2014, for a formal definition)
Computational run time (and complexity)	Gauge of computational costs
Robustness of tuning parameters	Should be explored from a Bayesian and a frequentist perspective.

Abbreviations (from table above)	
TP: true positives	True matches that are identified as such by the linking model.
FN: false negatives	True matches that are not identified as such by the linking model.
FP: false positives	Pairs that are not a match but are identified as such by the linking model
TN: true negatives	Pairs that are not a match and are not identified as such by the linking model.
FNR: false negative rate	
FPR: false positive rate	

Steorts went on to discuss a remaining question for entity resolution, namely, how do researchers make sure a method is robust? The robustness decision needs to be considered in light of the type of methods used to perform the entity resolution. A semi-supervised method needs to be robust to different choices of the training/test data and any tuning parameter(s). Probabilistic and Bayesian methods need to be robust to choices of hyper-parameters and/or tuning parameters. It is also important that the methods and models are robust to a number of

different datasets. In practice Steorts generally applies the method to 5-10 different data sets to make sure the method is fully robust from simple datasets to the most complicated datasets.

As a wrap-up, Steorts stated that, in order to continue growing, the field of entity resolution needs more workshops and forums, more open source software, work that promotes reproducibility, more evaluation metrics and comparisons, more transparency, and more ethical use of data. Transparency refers to the importance of academics and statistical agencies showing what they are producing and what they are doing, even when it isn't going well.

Discussant: William Winkler, U.S. Census Bureau

William Winkler's discussion of Steort's presentation raised additional considerations not fully covered in Steorts's presentation. First, he discussed the goal of combining datasets to create larger, cleaner sets of data for policy analysis. In creating such datasets, issues arise in cleaning up the original source files, creating the merged data files, and then adjusting statistical analysis for linkage error. He began with this in order to bring to the audience's attention the problem of additive error. If there is a 5% error in each of two linked data files and there is a 5% matching error, the resulting dataset has 15% error. Researchers (and data users) need to be aware of this error to determine what analysis is possible with the known level of error. Additionally, Winkler noted it is rare to have good test decks to evaluate how well the matching process is working. Overall, Winkler was very optimistic about Steorts' work, noting that it should improve Census's (and other's) methods.

The session concluded with a question and answer session. When asked what needs to be shared with the public in terms of metrics, Winkler responded that the two standard metrics, which are the same as those in computer science and in the health literature, are precision and recall. For record linkage, the reduction ratio is also a useful metric.

Session Two: Harmonization of Data across Sources

Chair: Linda Young, USDA National Agricultural Statistics Services

Speakers:

1. Ben Reist, U.S. Census Bureau. *Leveraging Survey Methods to Improve Administrative Record Estimates*
2. Don Jang, NORC. *Data Harmonization in Survey Data Integration*
3. Scott Holan, University of Missouri. *Recent Advances in Spatial and Spatio-Temporal Change of Support for Official Statistics*

Ben Reist, U.S. Census Bureau - "Leveraging Survey Methods to Improve Administrative Record Estimates"

Ben Reist summarized recent work on using survey estimates to assess the quality of administrative record (ADREC) data instead of the more common approach of using ADREC data to improve the quality of survey data.

Reist explains that data collected from surveys can be used to address coverage issues that are often present in ADREC sources. In some instances, survey data also can be used to address measurement error when the survey is considered to be of higher quality than the administrative record source.

Using an overlapping-frame framework, survey data collected from a more representative frame can be used to understand the coverage issues associated with the ADREC source. The formula for quantifying coverage error is as follows:

$$\hat{t} = \lambda \hat{t}_a^A + (1 - \lambda)\hat{t}_a^S + \hat{t}_s$$

The first two terms are the estimate for the population covered by the administrative data, which is a weighted average of the survey-based and the administrative-record-based estimate. The final term is the estimate for the population covered only by the survey data and that is what provides an estimate of the extent of coverage error in the administrative data.

With respect to measurement error, survey data can be used to correct errors in individual records and also to monitor the overall quality of the ADREC data source. A rigorous survey data collection effort can be treated as a gold standard and used to understand the bias properties of ADREC using the following formula from Lohr and Brick (2012):

$$\hat{\delta}_a = \sum_{i=1}^{n} w_i (y_i^S - y_i^A)$$

where w_i is the survey weight, y_i^S is the survey estimate and y_i^A is the ADREC report. The two formulas above can be combined to produce an adjusted ADREC estimate, combining coverage and measurement error adjustments:

$$\hat{t} = \lambda(\hat{t}_a^A + \sum_{i=1}^{n} w_i(y_i^S - y_i^A)) + (1-\lambda)\hat{t}_a^S + \hat{t}_s$$

The first term of this model can be considered a GREG estimate with an intercept of 0 and slope of 1 (i.e., basically a substitution).

Reist notes that treating survey data as the gold standard is a strong assumption. In the reverse situation, where the administrative record is treated as the gold standard for measurement, the formula for estimation is simplified and only addresses coverage.

Abowd and Stinson (2011) suggest a model for estimation with no assumption of a gold standard. The method involves applying a probability as to the likelihood that the survey or the ADREC is correct and using this probability to weight the impact of the source of the final estimate. Reist notes that this is a useful method for performing sensitivity analysis and understanding the impact of relying more or less heavily on one source or the other.

Reist notes that further refinements to the proposed model are possible, including extensions to incorporate multiple ADREC sources. He concluded by listing a number of open questions to address. How should nonresponse be addressed? How should the sample be allocated across the non-overlapping portion of the frame? How should this integrated data approach be done in a multivariate setting with multiple estimates of interest?

Don Jang, NORC - “Data Harmonization in Survey Data Integration”

Don Jang presented an overview of data harmonization and an example of harmonization in action in the Scientists and Engineers Statistical Data System (SESTAT), a study funded by the National Science Foundation (NSF) that leverages estimates from three surveys.

Data harmonization is ‘the process of mapping and synchronizing data derived from multiple sources into a coherent data file for analysis.’ There are a number of challenges associated with harmonizing data—data sources are often hard to link; data from different sources vary in cost, what they report and who they represent; and there are no universal data quality measures available to evaluate the harmonized data.

SESTAT is comprised of three surveys that collect nationally representative data on employment, education and demographics of U.S. scientists and engineers. Each survey (the National Survey of Recent College Graduates (NSRCG), the Survey of Doctorate Recipients (SDR), and the National Survey of College Graduates (NSCG)) samples from a different population, collects different data, and presents different quality issues.

By combining all three sources, SESTAT's cross-sectional coverage is broadened, producing a nationally representative sample of the US scientist and engineer population. Harmonization also increases temporal coverage, and allows for trend estimation and longitudinal analysis.

Harmonization is implemented at the question level by making sure naming conventions, formats, coding schemes and editing rules are coordinated and standardized across surveys. Response rate calculations also must be coordinated so that they can be compared and appropriately utilized for weighting.

Regarding imputation, there are 200 items across the three surveys and missing values for all of these items are imputed, primarily using hot-deck imputation. The variables used to determine the nearest neighbor may vary from one survey to another. To help standardize imputation, NSF has begun to provide suggestions about which variables to use. Similarly, weighting procedures are being standardized so that auxiliary variables and response propensity models are performed similarly across surveys when possible. Combined data products must be adjusted for multiplicity and different weights need to be developed for cross-sectional and longitudinal analyses. Each survey also has its own variance estimation method and data users therefore require guidance on how to perform estimation with the combined data.

Jang concludes by stating that data integration requires significant resources and that harmonization should be treated as a separate step in the data development process that should be built in from the beginning.

Scott Holan, University of Missouri - "Recent Advances in Spatial and Spatio-Temporal Change of Support for Official Statistics"

Scott Holan summarized his experience in using data with certain spatial and temporal properties to produce estimates with different spatial or temporal properties (for example, using 1-year or 5-year period estimates to produce 3-year estimates; or using estimates from two regions with overlapping but different borders to make inferences about one of the regions).

There are two approaches to producing estimates on multiple spatial scales (referred to as change in support, COS). The bottom-up approach relies on estimating variables on a finer resolution on the support source (the regions from which you have data) and then averaging these estimates across the target support (the region you wish to estimate). The second approach is top-down which involves partitioning the source and target support. The bottom-up approach is the focus of the talk.

Bradley, Wikle and Holan (2016) summarized a bottom-up method for accounting for spatial change of support for count-valued survey data. The method uses a Bayesian statistical model that incorporates dependencies between different regions and leverages survey variances to address the differential sampling errors that are introduced by the different support sources.

Holan describes an example of this methodology in which the Department of City Planning in New York City wanted to use American Community Survey data aggregated from census tracts to produce poverty estimates for community districts.

Next, Holan described a methodology for facilitating spatio-temporal COS using 1-, 3- and 5-year ACS estimates. Holan showed that 1- and 5-year estimates can be used to produce 3-year estimates with improved coverage, smaller standard deviations, and similar point estimates to the actual 3-year estimates produced by the ACS *(*Bradley, Wikle and Holan 2015).

Spatial estimation under COS may suffer from aggregation error, otherwise known as the ecological fallacy or Modifiable Areal Unit Problem (MAUP). To address this issue, the size of the error for potential source regions must be quantified in order to select the source that minimizes the error. A recent paper of Holan's attempted to address this regionalization problem; Bradley, Wikle and Holan (2017) presented the criterion for spatial aggregation error (CAGE) which compares the difference in variance between the fine-scale and aggregated-level estimate. CAGE allows for identifying optimal regionalization and evaluating the severity of MAUP in a given spatial domain.

Holan pointed out other areas of interest with respect to spatial and temporal estimation, including papers on combining data from multiple sources (Bradley, Holan and Wikle, 2016; Wang et al., 2011) and papers on combining data from different temporal sampling frequencies (Holan, Yang, Matteson and Wikle 2012; Porter, Holan, Wikle and Cressie 2014).

Session Three: Combining Data by Statistical Matching, Imputation and Modeling

Chair: Lisa Mirel, National Center for Health Statistics
Speaker: Jerry Reiter, Duke University.
Discussant: Ed Mulrow, NORC

Jerry Reiter, Department of Statistical Science, Duke University and U.S. Census Bureau – "Blending Data through Statistical Matching, Modeling, and Imputation"

Jerry Reiter's presentation focused on the different approaches and models of statistical matching. Statistical matching is used to blend data sets without unique identifiers and may be used to match datasets without overlapping observations. If the goal is to learn about the associations between Y and Z, but File One contains variables X and Y and File Two contains variables X and Z so that Y and Z are never observed together, the fundamental problem is that the joint distribution of Y and Z can never be estimated from the data alone. Some form of external information is needed to proceed with the statistical matching. This external information

could be assumptions made about the association between Y and Z given X, another dataset with Y and Z (and ideally X) observed simultaneously, or constraints on the associations drawn from other sources.

Even with external information, however, assumptions are still needed for statistical matching. The most common assumption is conditional independence, i.e., that Y is independent of Z given X. The methods typically used for statistical matching implicitly assume this, including methods such as nearest neighbor hot deck in which nearness is assessed based on the values of X and regression modeling in which X serves as the explanatory variable or vector. The nearest-neighbor hot deck method is the most commonly used method, but there are a number of pros, cons and quality concerns associated with it. The pros are that it is easy to explain to others, it is familiar to statistical agencies, and it can generate realistic multivariate imputations. The cons are that conditional independence is a strong assumption that is difficult to evaluate. If this assumption is not true, matching could be unreliable. Additional cons are that the researcher has to select a distance function, single imputations underestimate uncertainty, and the method can cause difficulties with later edits. The quality concerns include understanding how similarly the X variables on the two data sets are defined, how contemporaneous the data files are, and whether the underlying survey had a complex design (and what effect that has on matching different datasets). Additionally, researchers need to understand how to propagate uncertainty from the matching and how to conduct a sensitivity analysis.

Regression modeling can be used to estimate a model that predicts Y from X, which then can be used to impute the missing values of Y in the second file. Like the hot deck approach, regression modeling of this sort also has pros, cons, and quality concerns to consider. On the positive side, regression modeling is more flexible than hot deck (e.g., it can use predictive engines from machine learning), it can be used to specify models so that imputations satisfy edits, the researcher can check the quality of the regression model, it can be used as a prescriptive and flexible approach to sensitivity analysis, and it naturally leads to multiple imputation for uncertainty propagation. The cons are that the researcher still needs to make unverifiable assumptions about alpha (the level of conditional dependence between Y and Z), and also must select a model, leading to a risk of misspecification. The quality concerns are the same as those listed for the hot deck method above.

If the external data that are used to move forward with the statistical matching include another auxiliary dataset with Y and Z observed (ideally with X), this information can be used to reduce the reliance on the conditional independence (and other unverifiable) assumptions. If all the Y and Z variables are observed for all the X variables, then regression modeling can be used to impute missing values of Y (and Z) for the two files. If an arbitrary subset is observed in one (or multiple) file(s), where only some variables of Y and Z are observed jointly (possibly with some X variables), it is possible to append the auxiliary data to the other data set(s) and estimate a joint model using the incomplete data. It is important to construct the appended data so as not to distort the marginal distributions of X and Y, and of X and Z. (see Fosdick, De Yoreo, and Reiter 2016 for an example of this approach).

As with other approaches, however, there are pros, cons and quality concerns with using auxiliary datasets. The pros of using auxiliary information are that this reduces reliance on

unverifiable assumptions, the researcher can specify models so that imputations satisfy edits, the quality of the auxiliary data models can be checked for predicting marginal distributions of observed variables, and it naturally leads to multiple imputations for uncertainty propagation. The cons of using auxiliary data include the fact that the researcher still has to choose the model and make some unverifiable assumptions about the unobserved marginal and conditional relationships. Additionally, the researcher has to be careful about constructing the auxiliary data, especially when using joint models, and it can be difficult to do sensitivity analysis with flexible joint models.

Reiter wrapped up his presentation by discussing his thoughts on what agencies might report when doing such blending. Agencies performing statistical matching should provide the meta-data for files used in matching and should be transparent about the steps taken to harmonize X variables (e.g., X asked in similar ways?) and other edits that are done; the assumptions used in developing the models as well as the matching method and X variables used; the assessments of quality of fit of regression models; and the results of the sensitivity analysis. Additionally, agencies that use auxiliary data should be transparent about the potential selection biases in these data, the specification of conditional distributions in the auxiliary data, and the combinations of variables that were not observed jointly, which as Reiter noted is harder to report but still important.

Going forward with statistical matching research, Reiter listed a few questions that need to be tackled. There needs to be a better understanding of the usefulness of convenient, non-representative auxiliary data. There is also a question about how researchers should implement the joint distribution approach listed above, and once implemented, how users should be informed about what they can expect to estimate well and what they cannot for their specific queries. Finally, more needs to be understood about how uncertainty can be propagated in this context.

Discussant: Ed Mulrow, NORC

In his discussion, Ed Mulrow expanded the context of Reiter's work by discussing the overall purpose of combining data. Data are combined to increase either the length or the width of the original dataset in order to improve coverage (survey data from different frames), increase sample size, and/or bring together variables from different files. The actual data linkage can be done at the entity level (entity resolution and statistical matching/data fusion), the area level, or through a combination of entity to area (multi-level). Statistical matching, which Reiter focused on, can be used when the record's measurements are at the same level and there is little to no overlap of records across samples.

Session Four: Disclosure Avoidance: Frameworks, Techniques and Quality Issues

Chair: Shelly Martinez, Office of Management and Budget
Speaker: Latanya Sweeney, Harvard University

Discussant: John Abowd, U.S. Census Bureau

Latanya Sweeney, Harvard University - “The elusive sweet spots of privacy and utility”

Sweeney’s presentation focused on determining the optimal approach for protecting privacy while preserving data utility. The traditional belief system with respect to the relationship between privacy and utility is that there is always a tradeoff such that utility must be sacrificed to maintain privacy. Sweeney explained that successes in balancing utility and privacy occur when the ‘sweet spot’ is identified, such that the maximum possible level of utility is achieved while preserving privacy. There is still much work to do to achieve the sweet spot and little scientific evidence to inform the true re-identification risk inherent in survey data.

In 1997, Sweeney was surprised to find that she was able to re-identify the governor of Massachusetts in a source of data on health care utilization collected by the state’s Group Insurance Commission and released as a public use data file believed not to compromise any individual’s privacy. She was able to do this by merging the data with voter registration data available for purchase and matching based on the overlapping fields (zip code, birth data and sex) to uniquely identify the governor. Sweeney (2000) points out that in the 1990 census data, 87% of Americans are unique based on their date of birth, zip code and gender. This percentage drops as the data become more generalized, but it never truly goes to zero due to individuals such as park rangers who may be the only residents of a particular place or county. This research was cited in the preamble of the HIPAA Privacy Rule and other similar efforts around the world.

Sweeney suggests that improving disclosure prevention should be done in a similar way as data encryption, where people expose vulnerabilities in the current method in order to develop an improved new method. This process led to the strong encryption standards we see today and a similar process could be used as a means for reducing re-identification risk. The remainder of the talk focused on Sweeney’s research into identifying vulnerabilities in existing disclosure prevention methods with the goal of improving them.

Sweeney (2015) gathered anonymized medical data from Washington State and combined it with published accident reports from newspapers and found that 43% of the records could be re-identified. Only Washington and California changed their health data privacy laws as a result of this and Sweeney concluded that, in order to convince the other states to follow suit, someone would need to perform similar re-identification risk analyses in each state. When similar studies were performed in Maine and Vermont, each found re-identification rates around 30% (Yoo, Thaler, Sweeney and Zame (2017)).

Sweeney, Yoo, Perovich, Boronow, Brown and Brody (2017) examined the effectiveness of HIPAA Safe Harbor Laws which limit reporting on date of birth, gender and zip code. The authors used data from a health survey conducted by the Silent Spring Institute which also asked about housing information (number of bedrooms, bathrooms, square feet, and decade

built) and combined this with property ownership data and data provided from a data broker firm. Merging these data sources led to a 25% re-identification rate.

Sweeney discussed the efficacy of expert determination and best practices related to disclosure prevention. In a re-identification study of those who had taken the California bar exam, Sweeney, Von Lownfeldt and Perry (2017) found that four expert-developed guidelines for preserving privacy failed to prevent re-identification; these methods included techniques such as recoding, aggregating, duplication, dropping records, and replacing values with z-scores.

Regarding formal protection models, Sweeney explained that k-anonymity can eliminate data utility and methods such as differential privacy and using synthetic data often violate data delivery standards on accuracy.

Sweeney presented three recommendations stemming from the U.S. Commission on Evidence-Based Policymaking for what could be done to maximize utility while maintaining privacy: 1) Assess risk of a particular data set instead of applying a standard privacy prescription across all data sources, 2) Use a sliding scale of access to increase privacy protections as public availability increases, and 3) Be transparent about what disclosure prevention methods were applied.

Sweeney, Crosas, and Bar-Sinai (2015) suggest using the Data-tags System, which employs a six-tiered rating to specify the security level and access requirements for a particular data set.

Sweeney concluded by noting that there are a number of tools available to improve privacy. However, none of them constitute a panacea. Therefore, more work is needed to improve policy and technology to ensure privacy moving forward.

Discussant: John Abowd, U.S. Census Bureau

In his discussion, John Abowd echoed Sweeney's conclusion that continuous improvement is how data encryption became as strong as it is today and that data disclosure needs to go through the same level of scrutiny in order to reach an acceptable level. He noted, however, that there was much enthusiasm to improve data encryption to facilitate economic growth and stated that he is skeptical that the same enthusiasm will be present for limiting data disclosure.

Abowd noted that, in 2003, a significant event occurred in the data disclosure world, namely, the publication of the database reconstruction theorem. In their seminal paper, Dinur and Nissim (2003) proved that publishing a certain number of accurate statistics from a confidential database will allow for reconstruction of the confidential microdata "to an arbitrary level of accuracy."

One way to avoid the consequences of the database reconstruction theorem is to introduce random noise into the data that is statistically independent of any of the other distributions used to model the underlying data. Abowd notes that this is a necessary but not sufficient condition

for preventing disclosure. He believes this fact has been underappreciated by the statistical agencies in their efforts to improve disclosure prevention.

He believes further that the notion that we cannot share methods for maintaining disclosure for fear of causing disclosure is no longer acceptable, and that evolving the disclosure prevention field to maximize privacy and utility will require advances in technology.

The good news is that the principal mission of a statistical agency is to produce reliable data, meaning that the inference has the properties that you expect it to have, not necessarily that the data point in the microdata is exactly what was obtained from the respondent. Output noise infusion is the method that Abowd believes we should pursue to facilitate inference while protecting privacy.

Abowd suggests a formal privacy model for all publications in order to maximize the utility of a data source while maintaining privacy.

Workshop 3: Quality of Output Data

Workshop 3 was developed to help identify issues related to conveying information about the quality of the resulting output data and information drawn from integrated data. This workshop was divided into four sessions. The first two sessions each had a single presenter who focused on conveying quality to data users; the first session focused on survey data alone and the second on a large, integrated data project. The third session had two presenters who discussed the existing frameworks that have developed to assist with conveying quality information. The fourth session tied the current workshop to the previous two and included a discussion on the next steps the federal statistical agencies can take to identify and covey integrated data quality to the users of this data.

Session One: Break in Series

Chair: Richard Reeves, National Center for Education Statistics
Speaker: Lynn Langton, Bureau of Justice Statistics
Discussant: John Eltinge, U.S. Census Bureau

Lynn Langton, Bureau of Justice Statistics - "*Identifying and Addressing and Break (Blip) in Series"*

Lynn Langton presented the lessons learned from a redesign of the National Crime and Victimization Survey (NCVS). The redesign caused a break or, as Langton defined it, a "blip" in this longitudinal data series.

The NCVS began in 1973 and was designed to be a complement to the administrative crime records collected by the FBI, capturing crimes that are not reported to the police as well as those reported to the police. It is a nationally representative, interviewer-administered survey. The sample is a stratified, multistage cluster sample and includes persons aged 12 and older. Each selected household stays in the sample for seven consecutive waves of the survey. The first survey is conducted in person with the following waves intended to be conducted over the phone, though in reality, approximately 50% of the additional waves are conducted in person. The survey does not include the homeless, persons in institutional group quarters, and persons on military bases.

There were two driving factors for the redesign that occurred in 2016. The first was a routine need to update the sample to account for population shifts that occurred between the 2000 and 2010 Censuses. From 2006 to 2015, the sample design of the NCVS was based on the 2000 Census. In 2016 the sample was adjusted to reflect the 2010 Census. This new sample design will stay in place until 2026. The second driving factor was a desire to expand and reallocate the sample to enable state-level estimation for the 22 largest states; this was a shift away from

producing only national crime estimates to more detailed state-level estimates. Phase 1 of the redesign began on January 1, 2015 and the redesign was completed December 31, 2017.

There are a few things to note about the redesigned survey. In the 2016 data collection, there were three types of counties included in the sample. Continuing counties were those included in both the 2000 and 2010 Census designs; they tended to be large, self-representing counties and made up approximately 50% of all the included counties. New counties were new in the 2010 Census design. Outgoing counties were part of the 2000 design but would be phased out with the redesign. The sample size was also dramatically increased from 96,000 households in 2015 to 135,000 households in 2016. Of the 2016 households, 20% were interviewed by new interviewers, referred to as FRs, instead of the more usual 7% from 2015.

The redesigned sample included components that were expected to impact the NCVS estimates. First, earlier redesigns showed that “new” FRs affect the NCVS estimates. In the 2006 redesign, new areas added to the sample were predominantly rural and new FRs were hired to conduct those interviews. The violent crime rates obtained for those areas were 33% higher than the rates in the outgoing areas. Since the new areas were predominantly rural, the effects could be isolated. Second, previous training impacted the NCVS estimates. In 2011 there was a Refresher Training and FRs who participated in the training had higher numbers of crimes per interviewed person than the cohort of “untrained” FRs. Third, a potential new household effect was expected with more crimes reported in the households receiving the first of the seven waves of interviews. This effect may be explained through telescoping errors, with interviewed persons pulling in crimes that occurred before the start of the six month reference period. In addition, interviewed persons in their second through seventh wave may have developed an understanding of the types of incidents that are in scope for the survey and thus report fewer crimes, or may have become less willing to answer affirmatively to the crime screener questions, having come to realize that affirmative answers result in longer interviews. Fourth, there was a potential new area effect. The characteristics of the new areas could be associated with higher or lower crime than the outgoing areas.

The impacts of these effects were initially analyzed when the 2015 data were compared to half a year of the 2016 data. This analysis showed an increase in violent crime, but not an increase that was necessarily out of the realm of normal. Most of the victimization point estimates that were examined increased, but not all of these increases were significant. When the data were examined by level of interviewer experience and survey wave, it was found that, regardless of the year, there were higher rates of violence among new FRs (less than 6 months experience) than among experienced FRs (6 months experience or more). There were no significant changes in victimization numbers from 2015 to 2016 among experienced FRs, regardless of the wave. There were also higher rates of victimization for households in the first wave compared to households in the second through seventh wave, regardless of the year. Overall, new households and new FRs contributed to higher victimization rates in 2016.

Recognizing that the NCVS estimates had been impacted by the methodological choices associated with the redesign, decisions about whether and how to mitigate the effects were

necessary. After careful consideration of a number of adjustment options, Langton was not convinced that satisfactory adjustments could be made. Considering the fact that the new FRs and the new households were also more likely to be in new areas, it was possible that different household characteristics rather than having more new FRs and more new households might account for the differences in crime rates. Further, it was not obvious whether new FRs should be defined based on the length of time they had been an FR or the number of interviews they had conducted, creating ambiguity about the appropriate way to downweight new FRs. And there would have been real challenges in explaining to users of the data why the adjustments were needed at all. Since the 2016 data were collected by better performing FRs and from non-fatigued households, they could well be more accurate and there is no external source against which they could be benchmarked to determine whether or not that was the case. Additionally, it was expected that the distributions of the numbers of new FRs and new households would return to 'normal' levels by 2017, making it possible to compare the 2017 rates to the 2015 rates for change estimates, meaning that any distortion in the 2016 numbers would be a "blip," rather than a break, in series.

For all of these reasons, the data were released without adjustment. The annual victimization report focused on 2016 victimization patterns, and did not compare the 2016 estimates to the 2015 estimates. The impact of the redesign was explained through the presentation of rates for continuing counties vs. new/outgoing counties. There was no statistically significant change in rates in continuing counties.

The Bureau of Justice Statistics (BJS) is continuing to explore whether adjusted 2016 estimates can be released that are more comparable to the 2015 estimates. BJS also is considering what can be done in the future to mitigate the effects of future changes to the NCVS.

Discussant: John Eltinge, U.S. Census Bureau

In John Eltinge's discussion of Langton's presentation, he tied the lessons learned from identifying and addressing the blip in the NCVS estimates to identifying and addressing a break in an integrated data series. The issue with the NCVS redesign was a loss of comparability of results over time. With non-survey cases, this issue could occur with a major change in or loss of a data source. A change might include a change to the quality of the data source, to the definitions of the variables, to the production system used, or to the host site. Any such changes could impact the quality of the integrated dataset, causing issues with (sub)population coverage; established web scraping, record linkage, de-duplication, data fusion, or imputation procedures; and possibly with model fit. There are practices and methods that can be used to mitigate these issues, but internal or external changes in sources, methodology or practice can produce a "break in series". This break might be seen in a change to the mean or dispersion structure of the final data product.

If a "break in series" does occur, there are implications for the transparent reporting of the quality of the data. For the NCVS, there were careful diagnostics of the predominant features

underlying the new design as well as an evaluation of the potential adjustments. For an integrated data set, designing capture and integration methods to be robust against primary "break in series" risk factors requires complex trade-offs among the quality, risk, and cost profile components. These trade-offs increase the need for two-way communication with stakeholders in order to understand the risk tolerance of the data users.

Eltinge concluded his discussion by asking the audience for concrete case studies to help explore what is known about potential "breaks in series," prospective mitigation strategies that are used, and how stakeholder priorities and risk tolerances are assessed

In the discussion with the audience, Paul Biemer asked whether conducting a parallel survey during the transition would have been possible. This is something he has done himself and it allows for a comparison of estimates. Langton noted that a parallel survey would have been very costly. Biemer followed up by asking whether an evolutionary or step-wise approach could be used to essentially hide the noise in the survey during the transition so there isn't a blip. With this sort of approach, the cumulative changes can be seen but there is no sharp discontinuity from one period to the next. Langton responded that this approach might be worth looking into, but there are cost implications that might make it infeasible for the NCVS.

Katharine Abraham noted that the primary interest of many data users lies with seeing changes over time. Even if it can't be done perfectly, when there is a break or breaks in an official series, a research series that is as consistent as possible over time can have value. Abraham gave the example of the Consumer Price Index (CPI) research series. There have been many small changes over time in the methodology used to produce the official CPI. The research series is a historical reconstruction of what the CPI would have looked like had current methods been used in the past. Although someone outside of the Bureau of Labor Statistics could perhaps have produced such a series, the CPI staff are in the best position to do so and having a more consistent series has been helpful to the data user community.

Rolf Schmidt commented that changes to underlying data sources can affect estimates. Assuming there will be changes to the underlying data sources, it is important to attach a version number to the data sources used to produce hybrid estimates. Each version should be tied to a new benchmark and it should be clear which versions are comparable. Related to this, Richard Reeves closed the discussion with an example of a definitional change that occurred in the IPEDS data collection. There was a change in outcome measures due to the college's improved ability to track students who leave their institutions. Definitional changes occur in the underlying source data and this also is important to note for the data users.

Session Two: Combining Data from Disparate Sources

Chair: Paul Marck, U.S. Census Bureau
Speaker: Trivellore Raghunathan, University of Michigan
Discussant: William Bell, U.S. Census Bureau

Trivellore Raghunathan, University of Michigan - “Combining Information from Multiple Data Sources: Challenges and Opportunities”

Trivellore Raghunathan presented a case study of using survey data to improve the accuracy and representativeness of a rich non-survey source of information on health expenditures and outcomes. The goals of the research were to measure trends over the period from 1999 through 2012 in the prevalence of various health conditions as well as the costs and expenses associated with treating those conditions in adults age 65 and older covered exclusively by Medicare. Furthermore, the research aimed to understand the dynamics behind changes in costs over time, that is, to understand whether cost changes were attributable to changes in treatment costs or changes in condition prevalence.

Multiple survey data and non-survey data sources were incorporated into the analysis. Each source varied in its accessibility and the types of data it reported; access to some sources required agreeing to strict data-user agreements, including conducting analyses in restricted data centers.

Two primary sources were used in analysis—survey and claims data from the Medicare Current Beneficiary Survey (MCBS) and survey data from the National Health and Nutrition Examination Study (NHANES). MCBS survey data were used as a means to compare prevalence estimates produced using the claims and NHANES data to assess their quality. For some conditions there were no self-report data available from NHANES and MCBS; in these cases, other survey data sources were used to estimate prevalence.

The analysis was restricted to those whose healthcare was provided exclusively by Medicare. Propensity score weighting was used to account for the excluded portion of the Medicare population who were enrolled in HMOs and for whom claims data could not be obtained. Cost estimates were multiplied to agree with published national health expenditure.

Claims data were used to flag incidence of a given condition and these data were aggregated to produce prevalence estimates for a given year. This method was reasonably well suited to certain chronic conditions requiring frequent treatment. It was less well suited to capturing low-frequency ‘acute’ conditions that do not require continuing treatment (for example, the 2010 prevalence estimates for “ever having had a hip fracture” would be underestimated if the treatment for the condition only occurred in previous years). In addition, high-frequency conditions such as hypertension may be underreported because mention of those conditions is excluded from claims containing a large number of other, more serious conditions. To

compensate for these shortcomings, NHANES data were used to calibrate the claims estimates. In addition to collecting self-reports of health conditions, NHANES conducts physical examinations of respondents. Therefore, NHANES incorporates diagnosis data from previous years as well as data on undiagnosed health conditions. For conditions with no self-reports in NHANES, a measurement error model relating calibrated and uncalibrated claims data was used to impute prevalence.

Raghunathan showed that the resulting prevalence estimates for 2009 based on this calibrated claims approach were more consistent between the claims data and self-reported prevalence estimates from NHANES. With respect to trends over time, the calibrated claims align more closely with the NHANES data and reflect the increased prevalence rates that have been observed over time.

To estimate the cost attributable to each disease in each year, Raghunathan developed a model to estimate the cost of condition X by comparing the costs of everyone with condition X versus everyone without the condition after controlling for covariates. An individual's estimated costs for a given year would be the sum of the estimated costs for their conditions. Raghunathan shows that these cost estimates (adjusted based on the calibrated prevalence estimates described above) correlate closely with observed costs.

To tease out the effect of cost change versus prevalence change on expenditures, Raghunathan estimated costs while holding one factor or the other constant (cost change or overall prevalence). This analysis implies a much larger contribution to expenditures from cost increases versus prevalence increase, though the contributions of the two sources vary by type of condition.

Raghunathan pointed to a number of issues of comparability across the data sources used in this analysis. Claims data come from doctors' reports with unknown reporting structure while NHANES estimates come from self-reports with known reporting structure. Contributing surveys varied by mode, sponsor and questionnaire design, each of which could impact reporting behavior and participation rates. He concludes that mixing probability and non-probability based sources is an exciting area for research in the world of integrated and big data and presents the opportunity to leverage non-survey sources of unknown quality. It would be a mistake, however, to dismiss probability-based research altogether, as probability samples are crucial for validation and for assessing representativeness.

Discussant: William Bell, U.S. Census Bureau

Bell provided an overview of small area estimation (SAE) and how it works when combining multiple data sources, then related this to Raghunathan's work. He also remarked on assessing the quality of small area estimation efforts.

Bell first summarized the Fay-Herriot model (1979), which has served as the foundation for a great deal of small area estimation work. In this model, the estimate for y in small area i is

based on a regression model that is specified with area-specific predictor variables, a set of coefficients on those variables, and an additional random error term specific to area i. Restricted Maximum Likelihood (REML) or Bayesian treatments typically are applied in order to fit the model. Predictions are formed as a weighted average of the direct survey estimates and the regression prediction. Best Linear Unbiased Prediction (BLUP) leads to less weight being given to the direct survey estimate when the sampling variance for those estimates is high (because the sample is small), with the bulk of the weight in that case given to the regression prediction.

To produce Fay-Herriot small area estimates, a population must be defined and the small area estimator for a given *y* must be an unbiased predictor for which decent estimates of sampling error are available. Covariates must have a consistent relationship with *y* across areas. In order to pull in estimates from multiple sources, it is better to create separate models for each source as opposed to incorporating the additional sources into a single model. This is because different sources will measure the same variables differently and inevitably come up with slightly different estimates.

Bell noted that Raghunathan's research has several aspects that are similar to an SAE design. He had sources that define the estimation target (NHANES and MCBS data) and his analysis includes covariates. Regarding prediction, instead of using Restricted Maximum Likelihood (REML) or a Bayesian treatment to identify the BLUP, Raghunathan used multiple imputation for prediction, which differs from typical the SAE approach.

Bell pointed out some issues that can arise in the SAE approach. First, all available data sources may be biased with respect to a target population. Second, covariates may differ in their relationship to *y* across different areas/populations. Finally, there may only be poor estimates available of the sampling variances of the predictor.

Bell suggested that variance reduction could be a good metric for assessing output quality when using SAE methods. A 25% reduction in variance is a good threshold for determining whether an SAE model is worth undertaking.

SAE does not address bias in the primary data source used to define the target population, but it could provide a means of estimating bias introduced by additional data sources being considered as covariates for modeling the target population.

Session Three: Frameworks for Assessing Data Quality

Chair: Jennifer Parker, National Center for Health Statistics
Speaker:

1. Paul Biemer, RTI. *Assessing and Improving the Accuracy of Estimators from Blended Data*
2. John Czajka, Mathematica. *Transparency in the Reporting of Quality for Integrated Data: International Standards*

The third session of the day focused on frameworks—both currently available and under development—that can assist in conveying quality standards to data users. This session had both a domestic component and an international component, discussing the developments in many international statistical agencies.

Paul Biemer, RTI – *"Assessing and Improving the Accuracy of Estimators from Blended Data"*

Paul Biemer began his presentation by relating the work around reporting on the quality of integrated data to the Total Survey Error (TSE) model. Biemer found that by relabeling some of the error sources and tweaking the delineation of those errors so that they capture more of the errors are that are also in integrated datasets, the TSE model can be quite helpful. The focus of this presentation was on hybrid estimators, where a hybrid estimator is defined as an estimator based upon multiple datasets and used to estimate a population parameter. Biemer's discussion was restricted to hybrid estimators based on the combination of a survey data set and an administrative dataset.

There are a number of steps that need to be taken to form a hybrid estimator. At each step, there is the possibility of introducing error into the estimator. To start, a unified dataset needs to be created. Each of the datasets, the survey and the administrative dataset, comes with its own inherent error sources, which then go through an integration process that also can introduce errors. The dataset errors can be compounded or mitigated through the integration and estimation process. The cumulative errors are in the final unified dataset, which is used to form the hybrid estimator. Then the estimate is used to infer the population parameter, and further inferential errors can occur at that stage.

When creating a unified dataset, it is important to think about:

- What error sources are associated with the unified dataset?
- Which of these pose the greatest intrinsic risks to data accuracy?
- Among the hybrid estimators that might be constructed from the unified dataset, which estimator minimizes the total error risk?
- What are the major intrinsic and residual error risks associated with the hybrid estimator?

- Which of these error risks could be further mitigated to maximally increase the accuracy of the hybrid estimator?

These questions focus on risk instead of mean square error. This is because the components of the mean square error might not be known, but risk can be estimated. Having even some idea of where the most serious errors likely are will provide a path to exploring those errors. This motivates the need for a framework to help identify the major error sources for a dataset, as well as the errors that could occur in each stage of the estimation process. Errors introduced in each stage can be pertinent to the output and can be tracked through the development process. A total error framework can be specified for each stage of the process and error mitigation can occur at various stages of the process; for an example, see the three-stage framework of Reid, et al. (2017).

In many cases it suffices simply to describe the errors in the final output (e.g., for registers, frames and other datasets; point estimates; indices). A total error framework for a generic dataset can be applied to the unified dataset in the estimation process. This total error model can be written as:

$$\underbrace{\hat{Y} - X}_{\text{total error}} = \underbrace{(\hat{Y} - Y)}_{(\varepsilon_1 + \cdots + \varepsilon_6)} + \underbrace{(Y - X)}_{\varepsilon_7}$$

ε_1 = Selection error
ε_2 = Coverage error (over-, under-, duplication)
ε_3 = Missing data error
ε_4 = Content error
ε_5 = Data processing error
ε_6 = Model/estimation error
ε_7 = Specification error

where Y-hat is the hybrid estimator and X is the population parameter to be estimated.
As discussed above, these errors might more readily be identified through risk estimates. There are two types of error risk. The first is intrinsic risk, which is the risk that an error source possesses if no steps are taken to reduce the error. It is the error risk of “doing nothing.” The second is residual risk. Residual risk is the risk of error for a source that remains after mitigation strategies have been applied. A risk profile such as the ones below can be used to compare datasets and estimators. Table 1 compares the survey, administrative, and unified datasets.

Table 1: Intrinsic risk profile of survey dataset, administrative dataset, and unified dataset

Error Sources	Survey Dataset	Administrative Dataset	Unified Dataset
Specification	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Undercoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Overcoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Duplication	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Selection	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Content	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Missing Data	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)

Table 2: Intrinsic risk profile of survey estimator and hybrid estimator

Error Sources	Survey Estimator	Hybrid Estimator
Specification	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Undercoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Overcoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Duplication	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Sampling/Selection	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Measurement/Content	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Data Processing	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Nonresponse/Missing data	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Modeling/estimation	Risk level (1, 2, 3)	Risk level (1, 2, 3)

To determine the risk level to assign to each call, Biemer suggests that speculation or asking for opinions about how large these risks are may be best. Once the levels are assigned, the researcher will be able to identify which error source may cause the most risk in the unified dataset and then try to mitigate that risk. Table 2 extends this risk profile to the estimators.

Many of the components are the same between the two tables above. If an error exists in the dataset, it will appear in the estimator as well. When the survey estimator and the hybrid

estimator are compared, if there isn't an overall reduction in the error risk in the hybrid estimator, determined by averaging the errors within the columns, then it may not be worthwhile to integrate the datasets to begin with.

Overall, a total error framework decomposes total error so that key subcomponents can be identified and addressed. A unified risk framework facilitates comparison across individual and unified data sources. An error risk profile can provide insights regarding the quality implications of unified datasets. They can help with assessing the intrinsic risks by error source and determining whether the residual risk can be reduced by data unification.

John Czajka, Mathematica – *"Transparency in the Reporting of Quality for Integrated Data: International Standards"*

John Czajka began his presentation by listing the reasons why international standards may provide useful references for the federal statistical agencies to use when developing reporting standards for integrated datasets. For multiple reasons, administrative data systems are more developed internationally than in the United States. At least in Europe, there has been a more rapid decline in survey response rates than in the United States. International organizations have been particularly active in the development of standards and there has been a recent focus on the use of administrative records and Big Data for official statistics.

Within the European Union's statistical organizations, the key organizations doing this work are Eurostat and the European Statistical System. Eurostat is the statistical office of the European Union and is charged with the production of official statistics at the level of all Europe for the European Union. The European Statistical System (ESS) is a partnership between Eurostat and the statistical authorities of the member states. The ESS Committee is charged with providing "professional guidance to the ESS for developing, producing, and disseminating European statistics". The key documents developed for the European Union are as follows:

- European Statistics Code of Practice for the National and Community Statistical Authorities (2011)
- Quality Assurance Framework for the European Statistical System (2015)
- ESS Handbook for Quality Reports (2015)

The European Statistics Code of Practice for the National and Community Statistical Authorities delineates 15 principles that address:

- The institutional environment (principles 1-6)
  - Professional independence
  - Mandate for data collection
  - Adequacy of resources
  - Commitment to quality
  - Statistical confidentiality
  - Impartiality and objectivity

- Statistical processes (principles 7-10)
  - Sound methodology
  - Appropriate statistical procedures
  - Non-excessive burden on respondents
  - Cost effectiveness
- Statistical output (principles 11-15)
  - Relevance
  - Accuracy and reliability
  - Timeliness and punctuality
  - Coherence and comparability
  - Accessibility and clarity

With some variation, the principles for statistical output are often included in frameworks of individual countries inside and outside Europe and are often described as “dimensions of quality.” For each principle, the Code lists several indicators of compliance. These describe actions that conform to the principle. For example, under accuracy and reliability:

> 12.1 Source data, intermediate results and statistical outputs are regularly assessed and validated
> 12.2 Sampling errors and non-sampling errors are measured and systematically documented according to the European standards
> 12.3 Revisions are regularly analyzed in order to improve statistical processes

The Code does not actually discuss these indicators. That is left to the ESS Quality Assurance Framework.

The Quality Assurance Framework for the European Statistical System is a framework produced to assist national statistical organizations in implementing the Code of Practice. It was designed as an aid in achieving quality, not for measuring or reporting it. It provides series of methods at both the institutional and product/process levels to facilitate achievement of the goal expressed in an indicator. For example, at the product/process level three methods for indicator 12.2 are:

- Periodic quality reporting on accuracy is in place
- Quality reporting on accuracy is guided by ESS recommendations
- Methods and tools for preventing and reducing sampling and non-sampling errors are in place

The purpose of the ESS Handbook for Quality Reports is “to provide guidelines for the preparation of comprehensive quality reports for a full range of statistical processes and their outputs”. The specific objectives of these guidelines are:

- To promote harmonized quality reporting across statistical processes and their outputs within a Member State and hence to facilitate comparisons across processes and outputs
- To promote harmonized quality reporting for similar statistical processes and outputs across Member States and hence to facilitate comparisons across countries
- To ensure that reports include all the information required to facilitate identification of statistical process and output quality problems and potential improvements

The Handbook also includes an appendix titled the “ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators.”

Czajka noted that the comprehensive quality reports addressed by the Handbook resemble U.S. quality profiles. A survey quality profile summarizes what is known about the sources and magnitudes of errors in a survey (Kasprzyk and Kalton 2001). It is a systematic and comprehensive review across the spectrum of survey activities in which both qualitative and quantitative results are brought together to allow an assessment of the quality of the survey operations and the data. The principles of relevance, timeliness, and accessibility are dimensions of quality generally not treated in U.S. quality profiles. Overall, U.S. quality profiles were produced for several federal surveys, but generally not more than once. They generally stopped being produced because they are resource intensive and staff to produce them were not available, they required information that did not exist, and their value to the survey producer was questionable.

It is important to note that the ESS Quality Assurance Framework and Handbook do not purport to be directed at integrated data but acknowledge that some of the estimates produced by European nations may be based on integrated data. Further, under the accuracy dimension there are separate discussions of statistical processes using administrative sources and statistical processes involving multiple data sources.

The Handbook states that when processes involve multiple data sources, a quality report should include how the process is organized, the individual segments that are included, and a summary of the quality aspects. The only suggestion regarding an assessment of the quality of the final product applies when a preliminary estimate is followed by a revision. The magnitude of the revision may be indicative of quality. A general recommendation is that whenever multiple data sources are used, a separate quality report should be produced for each data source and not just the combination of multiple data sources.

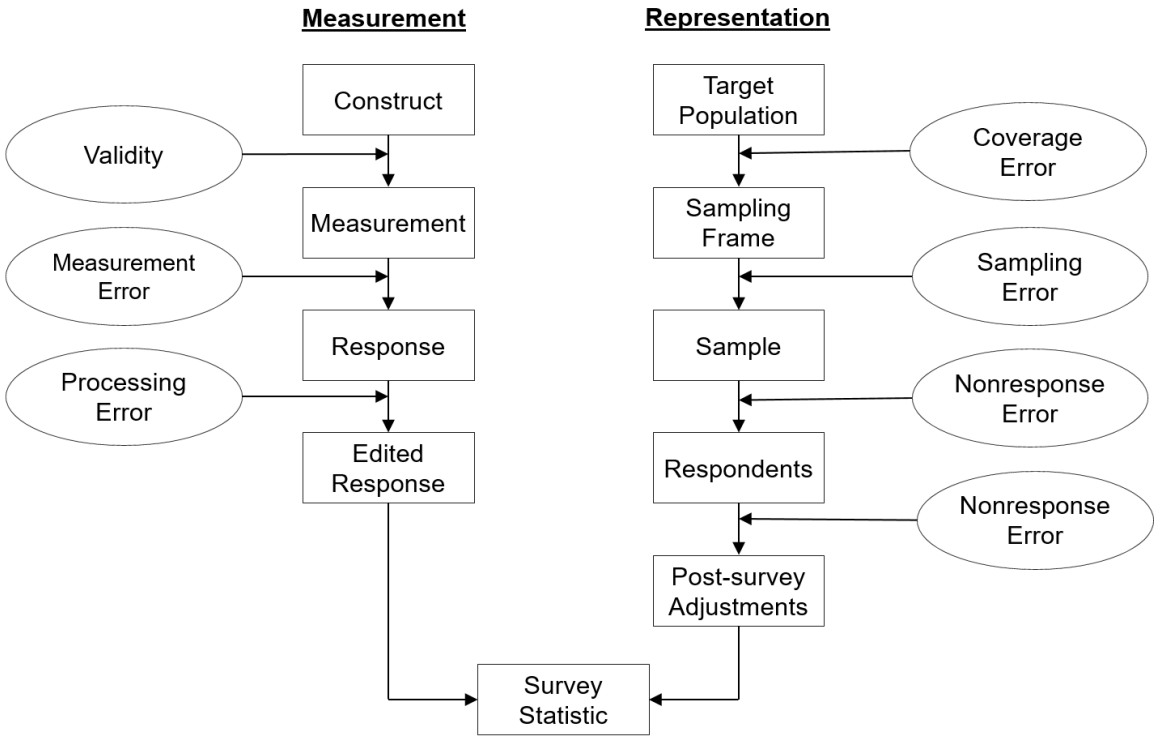
For the other quality dimensions not usually treated in U.S. quality profiles, relevance is focused on users of the statistical outputs and to what extent the data satisfy their needs. Different groups of users may have different needs. The one quality and performance indicator for relevance is the data completeness rate—the ratio of data cells provided to cells required. For timeliness and punctuality, the quality and performance indicators include the time lag between the end of reference period and initial results or final results and the time lag between the

announced target date and delivery of data. User feedback is identified as the best source of information for addressing accessibility and clarity. Quality indicators include how often users consult tables and metadata and the degree of completeness of the latter.

Other quality dimensions to note include coherence and comparability, cost, burden, and confidentiality. Coherence and comparability are assigned a high importance, with extensive information requested for the quality report. Quality and performance indicators address only “mirror flows” (inflows and outflows that should match) and length of unbroken time series. For cost, a quality report should include cost breakdowns by major components, although the difficulty of obtaining this information is noted. For burden, a quality report should include respondent burden in financial terms or hours, targets for reducing burden, recent efforts to reduce burden, and whether information collected is limited to what is absolutely necessary and cannot be obtained elsewhere. For confidentiality, there is a distinction made between legal requirements and data treatment.

While the ESS Quality Assurance Framework and Handbook do not include integrated data, efforts have been made to extend the TSE framework to integrated data in the international setting. Li-Chun Zhang of Statistics Norway has proposed a framework for integrated data based on the life cycle model of TSE in Groves et al. (2009). Statistics New Zealand (NZ) has adopted this framework as the basis for its own quality framework for integrated data.
The TSE model follows the life cycle of a survey from conception to the production of a survey statistic. The model builds on the idea that a sample survey consists of questions administered to a sample drawn to represent a target population. The model traces the dimensions of measurement and representation from an abstract construct and a target population through the design and implementation of a survey, culminating in a survey statistic. Error may be introduced at each stage as depicted in Figure 1 below.

Figure 1: Groves et al. (2009) life cycle model of TSE

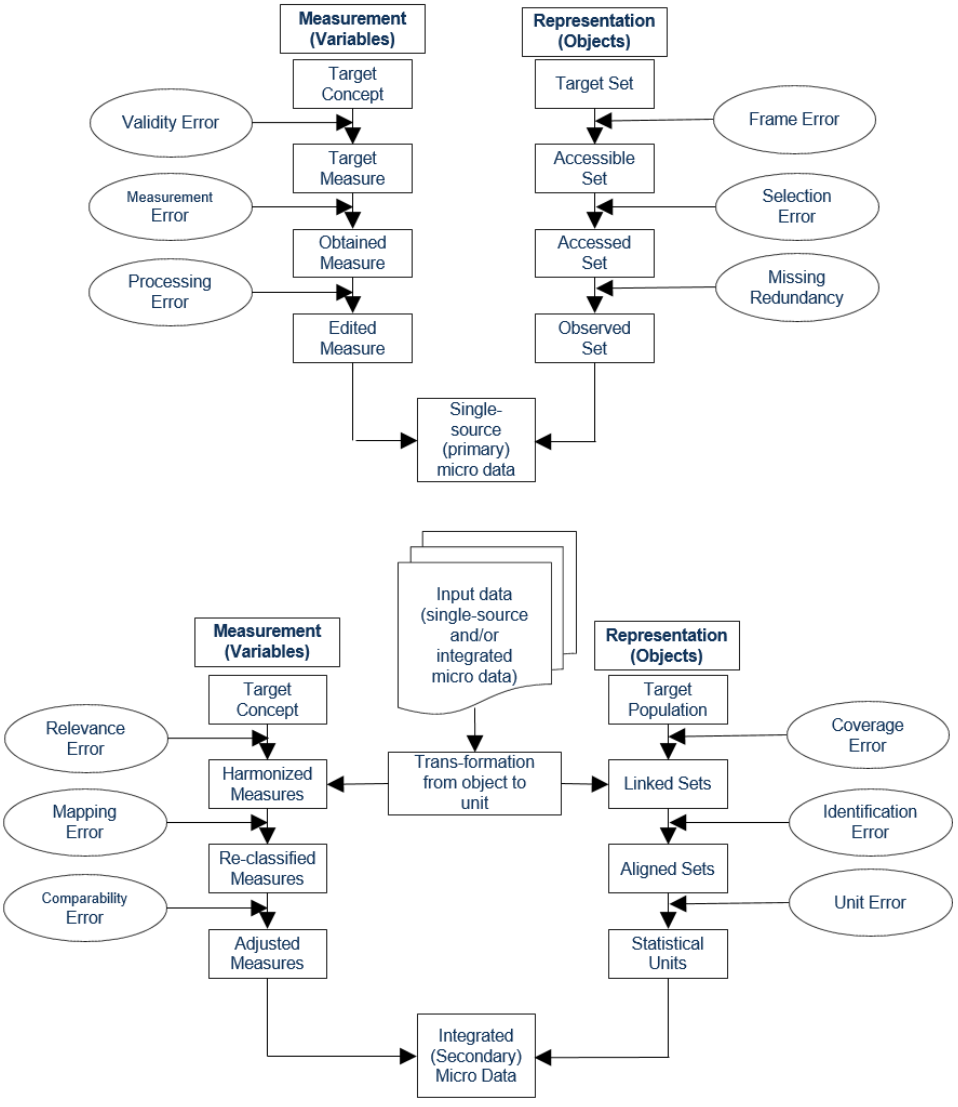


Source: Groves et al. (2009).

In Zhang's extension to a two-phase model, the end result of each phase is a micro dataset – not a single statistic. In addition, most of the concepts from Groves et al. have been renamed to accommodate the inclusion of data from administrative sources. For example, "measure" is used in place of "responses" and "sets" in place of "sample" and "respondents".

As can be seen in figure 2 below, phase one describes a single microdata source, but each input to the integrated microdata has its own phase-one assessment. Phase two shows the multiple inputs and depicts the sources of error for the integrated microdata. Harmonization on the measurement side and linkage on the representative side are critical steps in phase two. On the representation side, Zhang uses "objects" in phase one and "units" in phase two. For example, an input data source might collect information on jobs, but what is desired in the end is data on persons. The transformation of objects into units is shown in a box in phase two below the input of multiple data sets. Units themselves may have to be combined in some way – for example, persons may be aggregated to households.

Figure 2: Zhang proposed framework for integrated data based on the life cycle model of TSE



Source: Zhang (2012).

Zhang's conceptualization envisions an ideal target integrated dataset – the analog to an error-free survey statistic (at the bottom of the Groves model). Discrepancies between the target dataset and the final integrated dataset are analogous to the concept of total survey error in Groves et al. (2009). To assess the accuracy of the final dataset, Zhang develops the concept of empirical equivalence. Two datasets are empirically equivalent if they generate identical inferences; this does not require micro-level equivalence (e.g., same people). Zhang extends empirical equivalence to the assessment of public use data, where error is introduced to protect confidentiality.

With a mandate to make administrative data the data source of choice, Stats NZ faces the need to "assess and explain the quality of statistics that use multiple sources, including administrative data" (Holmberg and Bycroft 2017). Stats NZ issued a "Guide to Reporting on Administrative Data Quality", which uses Zhang's framework. The Guide includes quality indicators for each of

the phase one and phase two error sources. To extend the model further, Reid et al (2017) added a third phase of assessing the quality of final outputs – that is, the statistical estimates derived from the integrated microdata that are the endpoint of phase two. Quality indicators do not yet exist for phase three.

With the recent focus on the use of administrative records and Big Data for official statistics, Daas et al. (2011) of Statistics Netherlands present quality indicators for administrative data used as an input to official statistics. These indicators address five dimensions: technical checks, accuracy, completeness, integrability, and time-related factors. The dimension of integrability bears most directly on the integration of multiple sources. Four indicators are intended to capture how well the data source can be integrated into the statistical production system of an organization:

- Objects:
  - Similarity of objects in source with those used by organization
  - Ability to align objects in source with those of organization
- Variables
  - Usefulness of linking variables in source
  - Closeness of variables in source with those in other sources used by the organization

The indicators relate to phase one of Zhang (2012) in corresponding to objects (representation) versus variables (measurement).

Reflecting the focus on Big Data for official statistics, the U.N. created a Global Working Group on Big Data that is working towards standards. There are multiple teams addressing different aspects. The Big Data Quality Task Team published “A Suggested Framework for the Quality of Big Data” (2014). This report stated that “the application of either traditional data quality frameworks or those designed for administrative data would be an inadequate response to Big Data”. The Big Data quality framework has 11 dimensions nested within the hyper-dimensions of Source, Metadata, and Data which are applied to the phases of input and output. Possible indicators are listed for each dimension. This framework is in the early stages of development and is clearly a work in progress.

In summary, the international community has been actively developing quality reporting standards and frameworks that can be applied to survey data and extended to integrated data, administrative data, and Big Data. While there are many unique components to this work, there is a resemblance to the quality profiles that used to be produced for survey products in the United States. When considering how the United States could move forward with identifying and reporting quality standards for integrated data products, one solution may be to revive and extend the idea of a quality profile. Czajka noted, however, that the reasons these profiles stopped being produced are important to address.

Session Four: Workshop Summary

Chair: Linda Young, USDA National Agricultural Statistics Service
Speaker: Frauke Kreuter, JPSM University of Maryland

Frauke Kreuter, Concluding Remarks & Discussion. (JPSM)

In a summary of lessons learned across all three data integration quality workshops, Kreuter began by emphasizing the importance of assessing quality of integrated data products at the level of the estimate as opposed to the entire dataset. The success stories reported during the workshop typically have been with respect to one estimate. She pointed out that nonresponse does not necessarily lead to nonresponse bias across every estimate. Similarly, other quality indicators cannot be applied uniformly across an entire dataset.

Kreuter believes that, in order to move forward in an integrated data environment, existing research paradigms and ways of doing business should be re-examined. She suggested that the statistical community should get more comfortable with using proxies as data sources, recognizing that no data source is perfect and therefore some errors may be tolerable depending on the research question.

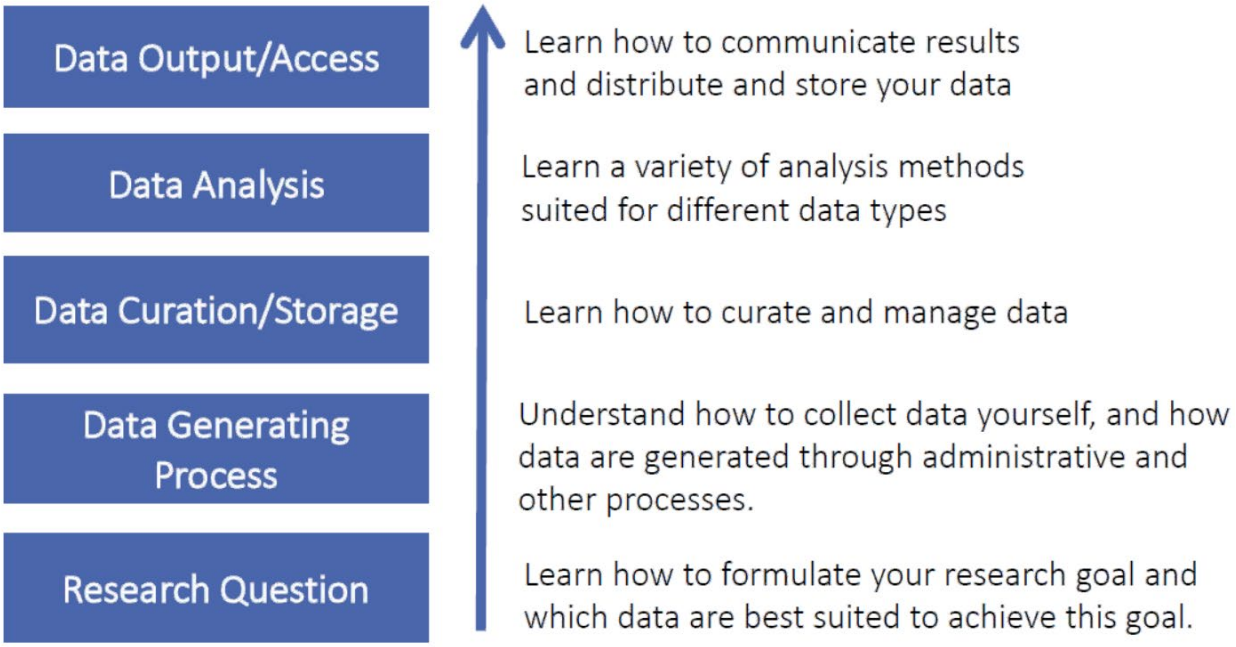
She noted an example from John Czajka's presentation about research requiring estimates to be at the household-level and not at the person-level—a restriction that can preclude the use of certain sources and limit analysis. Researchers should consider carefully what types of questions need to be answered with household-level data and what types of questions could be answered with more accessible person-level data.

Kreuter emphasized the importance of collaboration as a means of sharing knowledge and building on the efforts of others to reduce the burden of managing multiple data sources. She pointed out that, in the new integrated data environment, that burden will move from the front end of the process (data collection) to the back end (data processing and harmonization). Kreuter pointed to a recently published article (Oberski, Kirchner, Eckman and Kreuter, 2017) that presents a "generalized multitrait-multimethod modeling" approach for assessing the quality of integrated data. She noted, however, that using such a method does not produce a single metric for determining quality and instead produces indicators of quality specific to the research design at hand. An implication is that there may be considerable duplication of effort as the quality metrics of interest to calculate will differ for every research question. Collaboration can help to reduce the burden of developing research-specific quality metrics every time new research is conducted.

Kreuter questioned whether large survey datasets are the appropriate method of data production in the new environment. Such datasets may be less useful than using administrative records to sample individuals with a characteristic of interest for smaller data collection efforts. Kreuter noted the utility of smartphones for collecting data both passively and through surveys.

The coverage issues presented by using smartphones for data collection are diminishing as the bulk of the population in developed countries now owns one, though according to PASS data from Germany, prevalence rates still range from almost 100% for young people to less than 20% for those 80 and older.

Kreuter suggests that having an interdisciplinary team can be of particular value when working with integrated data products and measuring their quality. People from different backgrounds and with different knowledge of the various data products can work together to monitor quality and inform fitness for use based on the research question at hand. Such a team should be able to address the five stages in the analysis process outlined by Japec et al. (2015), summarized below.



Source: Usher in Japec et al 2015

Kreuter highlighted the Coleridge Initiative as a multi-university learning environment where researchers from a variety of agencies and technical backgrounds can collaborate and access a host of related administrative data sources in a secure environment, and then work together to address important research questions. Typically, these researchers would be restricted to accessing only the data generated by their own agency, limiting the utility of those data. By bringing different agencies and their data together in one computing and collaboration environment, the utility of each data product is increased and better research questions can be formulated based on input from multiple perspectives.

Following Kreuter's presentation, the topic of assessing quality at the estimate level versus data set took up most of the discussion. Frauke reiterated that the successful examples of quality

assessment involve a single estimate and that quality of an entire dataset is more challenging to quantify.

John Eltinge pointed out that datasets are often produced for a primary purpose and then leveraged for additional purposes; while it is impossible to assess quality for every possible purpose, it is important to have a framework and common language for assessing quality for a given data product so that users with different research objectives can understand its fitness for use. Nancy Potok also indicated the need for some general principles to adopt regarding quality and transparency that apply to all agencies so that the stakeholders can understand what was done to generate the data they're using (using models such as TSE). Nancy suggested trying to address integration issues (such as merging sources that measure the same construct but in different ways) at the onset of an analysis to get out ahead of quality issues that may arise later when sources are combined. Such an initiative would represent a long-term approach to facilitating data integration moving forward.

Raghunathan suggested that being transparent about assumptions underlying data processing decisions allows for others to question and test those assumptions to understand whether or not they invalidate the data for a given purpose.

Federal Committee on Statistical Methodology Conference

At the 2018 FCSM Conference held March 7-9, two sessions were dedicated to the workshops hosted by Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources and the Washington Statistical Society. The first provided an overview of the workshop presentations described above. The second was structured as a discussion with the audience aimed at identifying research priorities for enhancing the quality of integrated data and communicating important information about the quality of integrated data to potential data users. A summary of the second session appears below.

FCSM 2018 Session 2 Presentations

Each of the four presenters in this session—Chris Chapman (National Center for Education Statistics), John Czajka (Mathematica Policy Research), Joe Schafer (U.S. Census Bureau), and Linda Young (USDA National Agricultural Statistics Services)—were members of the FCSM Working Group/WSS workshop organizing committees. Alexandra Brown (Joint Program in Survey Methodology, University of Maryland) chaired the session and facilitated the conversation with the audience that followed the four presentations.

Chapman began the session with an overview of the first FCSM Working Group/WSS workshop, which focused on input data quality reporting. The central issue considered in the first workshop was what information should be provided to end users so that they can evaluate data quality and determine whether data are a good fit for their projects. Although there are a number of issues associated with measuring and evaluating the quality of data from sample surveys, there is still a generally agreed-upon set of metrics to facilitate consistent and relatively transparent reporting about the quality of these data. An important point emerging from the first workshop is that federal agencies and other data producers currently lack a clear standard to guide the reporting of quality metrics for non-survey data.

The presentations given during the first workshop suggest that, in many cases, producers of non-survey data are reporting quality metrics similar to those identified in data quality frameworks developed for sample survey data, such as the Total Survey Error (TSE) framework. For example, presenters talking about different sources of non-survey data discussed the purpose of the original collection, accessibility, accuracy, and respondent burden. The question remains, however, as to whether there are unique aspects of the quality of non-survey data that lack corollaries in survey data quality frameworks. If so and new reporting standards are needed, it also will be important to ask whether the same standards are applicable to both semi-structured and unstructured sources.

There are a number of issues still to be resolved before making suggestions about what should be reported to end-users to enable them to judge the suitability of an integrated data product for their needs. A central challenge will be dealing with the expanded variety of sources of input

data, including outside vendors who did not originally generate the data for research purposes. What should be done when data from a particular source lack what is considered by the statistical agencies to be standard documentation? How should this be reported to end users? Lack of documentation is often a problem with private-sector data resources, as the holders of those data resources may be reluctant to share information for proprietary reasons. In other cases involving data that were not originally collected for statistical purposes, data quality metrics of the sort statisticians are used to seeing simply may not exist. How should such data be evaluated and, more importantly, how can the quality of the data be conveyed to the public when the information about quality one would like is lacking?

Czajka opened his presentation by asking the audience to consider two examples that highlight issues in reporting on data quality:

1. Administrative data substituted for respondent data for a subset of variables in a sample survey
2. Two or more administrative datasets combined to produce estimates for a population that neither dataset covers fully

What kind of quality metrics might be constructed to report on the accuracy of the output in these two cases?

In the first example, the administrative data might measure exactly the same thing as the survey variables they replaced. Alternatively, the administrative data might provide a biased measure of the survey variable concept. In the first case, the same quality metrics that would be used to describe the survey responses can be applied to the administrative data. In the second case, the bias may need to be addressed in some other way.

In the second example, when administrative datasets are combined to produce an estimate for a larger universe than any one dataset represents, coverage is an important quality metric. How can coverage be measured if the target universe is not captured in any other dataset?

Schafer then steered the conversation to data processing, the topic of the second workshop. Data processing encompasses all of the steps taken between ingesting inputs from multiple sources (e.g., surveys, lists, purchased data, scraped data, sensors) and releasing the final products (e.g., estimates, analytic reports, actual or synthetic microdata). With such a large number and variety of topic areas, many of which are new and emerging, understanding how to assess and maintain quality when processing integrated data must be a team effort.

Schafer suggested two key areas as priorities for future research: 1) combining data from multiple sources; and 2) the role of modeling.

A recent paper by Lohr and Raghunathan (2017) reviews the basic techniques available for use in combining multiple datasets. Two well-known existing methods that have proven to be useful are post-stratification and calibration to external sources. Other methods that may be less well known include:

- Deterministic or probabilistic microdata record linkage
- Statistical matching, data fusion, or imputation—gluing together multiple microdata sets that estimate different aspects of a joint distribution and solving for the missing components
- Small area estimation—model-based smoothing and prediction of estimates for domains with few or no sample units
- Hierarchical modeling of estimates from different studies (as in meta-analysis), possibly with extra terms for study-specific bias
- Quasi-randomization inference and generalized weights for non-probability samples

With respect to the role of modeling in data processing, many of the current methods are based on formal models for:

- Population estimates of interest
- Unknown parameters (i.e., of one or more prior distributions), some of which may be inestimable from data at hand
- Inference under quasi-randomized designs, which require inclusion probabilities (for non-probability samples)
- Mode effects on measurement

Overall, due to the processing required to create an integrated dataset, the dependence on models and sensitivity to modeling assumptions tends to be greater for integrated data analyses than for analyses based on classical survey data. This underscores the importance of giving careful consideration to the specification and checking of models using integrated data. Diagnostics developed to check for model failure can be used, but sometimes these diagnostics do not provide direct evidence of model failure coming from the data itself. In such a case, sensitivity analyses that vary parameters that cannot be estimated are important, such as with extrapolation (i.e., identification of regions that are not measured well to the model). There is an increasing need to develop survey methodologists who are skilled in the science and art of modeling.

Young's overview of the third workshop began with a reflection on measures of uncertainty. Young highlighted this through a discussion of breaks in series, combining disparate data, and sensitivity analyses. Reflecting back to Lynn Langton's presentation in the third workshop, breaks in series may be reported when simple changes in methodology are made. Examples of such changes might include changes in question ordering, questionnaire design, or level of interviewer training. In truth, however, the effects on the estimates attributable to these changes may not be especially large relative to the underlying uncertainty in the estimates. This raises the question of whether there is a better way to communicate the underlying estimate uncertainty to users of the data.

As an example, Young displayed an image of combined satellite pictures of farmland in America. This combined image presented a fair amount of uncertainty in terms of where crops are grown. To address some of this uncertainty, statistical tools are used to improve the accuracy of the image, but there is likely to be residual measurement error due, for example, to

mismeasurement of spectral radiance or model misspecification. The challenge is to accurately reflect the uncertainty in the estimates based on the integrated satellite and survey data and to do so in a principled manner.

Young suggested that, when quantifying the uncertainty in estimates based on a data set that integrates survey and administrative information, the uncertainty should be discussed in terms of the survey data, the administrative data, and the integration and estimation process. Reported measures should reflect all of the uncertainty, but often they do not. Errors that are variable create additional complications for a data producer seeking to provide informative quality metrics. Additionally, standards need to be developed for reporting on the results of sensitivity analyses. In some ways, the uncertainty in an estimate and its sensitivity to model specification ought to be related, and there should be a way to communicate both to data users.

Audience Reactions to Session Presentations

Following the four brief presentations, there was a wide-ranging conversation between the audience and the speakers. Among the topics covered were the role of substantive experts, the household files that will be used to build the 2020 Census frame, tackling resource issues, the importance of setting quality standards for both datasets and estimates, how to train staff for future needs, and how to formalize a future research agenda. Some of the comments made during the conversation are summarized below.

The discussion began with a question about the role that substantive experts have in working with statisticians. For example, especially with modeling, there often is a need for knowledge about the data that survey methodologists or statisticians won't necessarily have. Schafer agreed that substantive experts often play an important role in developing appropriate models. This was clear in some of the case studies that were presented during the FCSM Working Group/WSS workshops. For example, there was a substantial role for substantive experts in the work of Trivellore Raghunathan (third workshop). Czajka noted that, in thinking about using administrative data as a substitute for survey responses, substantive expertise is needed to determine whether an administrative data variable really captures the same thing as the survey variable it is replacing.

The next audience comment related to the resources needed to tackle the problems identified during the workshops, especially in an environment in which data users do not know how hard the problems are or simply do not care. The responses to this comment focused on the importance of setting priorities. The question is whether the statistical agencies are prepared to make the handling of integrated data and the measurement of uncertainty in estimates produced using integrated data the focus of their resources. Agencies will try to reallocate resources strategically to methodological issues, but one of the challenges is making sure key stakeholders understand why this needs to occur. This tied into a discussion around the importance of having a research agenda that reflects the priorities of the Federal Statistical System as a whole. The statistical agencies have limited money and staff, and the same is true of the academic community. This means that it is critical to have a research agenda that

addresses the issues that are most important to the system as a whole, rather than a research agenda that is shaped too much by individual interests but does not move the whole system forward. Part of the motivation for the FCSM Working Group/WSS workshops was to demonstrate to stakeholders the work that most needs to be done.

The conversation then turned to education and training for the federal statistical community in the use of model-based approaches. As the agencies shift to model-based approaches (e.g., more reliance on imputation), there is also a question about how they maintain the public's faith in their estimates.

Chapman responded that the FCSM began some work on these issues about two years ago, with an effort to develop an agenda for ensuring that the staff of the statistical agencies have the skills they need as the agencies move into the realm of non-survey data. A successful agenda is likely to involve both training for current staff and hiring of individuals with technical skills who also have good communication skills. Additional responses focused on the need to hire people who want to be life-long learners, to preserve training budgets so that employees have the opportunity to acquire new skills, and to expand the search for good employees outside of the usual fields (for example, searching for biostatisticians and epidemiologists who have good modeling skills). Overall, statistical agencies want to move as rapidly as they can while ensuring the quality of the statistics they produce. If the statistical agencies were to become so eager to move down the path of using integrated data that incorrect estimates were published, that would invite difficult questions from the public.

Discussion of Areas for Future Focus

The next part of the session was a discussion of some of the questions highlighted during the three FCSM Working Group/WSS workshops. This discussion focused on disclosure risk, technological changes, and transparency around work that isn't going well.

Brown opened the conversation by asking the audience to respond to two questions associated with disclosure risk. Across the three workshops, particularly the first and third, there was considerable discussion of the need to protect the identity of the individuals whose information is contained in integrated datasets. Are there projects underway that can shed new light on thwarting disclosure risk? How is disclosure risk presented to data users?

A representative from the Department of Transportation (DOT), began the conversation by noting that this can become an issue when users working with economic data want more disaggregated information. For example, a user may want commodity flow data down to the county level, but providing this information might expose a single major employer in a particular county. In other words, disclosure risk isn't just a problem of record matching. In fact, if an agency has developed good models, it might be possible to reverse engineer some of the datasets that are out there and create a disclosure issue. We need to figure out how to deal with disclosure versus the desire for granularity. This may mean taking some fresh approaches to this tradeoff and producing a very different generation of products, for example, providing a slice from a matrix of estimates rather than trying to provide users with all of the individual cells.

Czajka noted another issue—whether integrating auxiliary data into surveys might itself violate laws about releasing the data.

A representative from the National Center for Health Statistics (NCHS), noted the active Research Data Center (RDC) program at NCHS. NCHS staff members having the opportunity to comment on the quality of proposals and every single proposal is reviewed for disclosure risk. While no standard quality framework yet exists and the disclosure review process is somewhat ad hoc, RDCs offer a means to let other people work with NCHS datasets and to do it under controlled conditions that are designed to prevent the release of sensitive information.

Brown then moved the conversation to questions regarding technological changes and the effect of such changes on data collection and data integration that came up in the first of the FCSM Working Group/WSS workshops. Are there examples of technological changes affecting or changing agencies' integrated data products? Are there examples of relying on a new technology or data source and then having the technology change or become proprietary or the data source cease to be available? Other issues related to technological changes include how to integrate new sources of data, what to do with existing sources, and how to communicate to the end user that the product is changing.

The first technology that was discussed was sensor data. Sensor data provide a lot of information about a very narrow slice of the world. In fact, the slice can be so narrow that, unless care is taken, a distorted picture may be obtained. For example, at the DOT, auto fuel economy is collected and analyzed on a moment by moment basis, capturing the effects of vehicle weight, speed and slope. One study found that very heavy trucks moving at very high speeds got great fuel economy on flat ground. It turned out, however, that the flat ground where the measurements were taken was at the bottom of a hill and it was necessary to check upstream to see this. Things like this reinforce the role of domain expertise. A person who is knowledgeable about the subject and has a good sense of what the sensor data should be showing can ask the appropriate questions if the patterns in the data do not seem right.

Turning to the health field, the point was made that when thinking about sensors and incorporating sensor data, it may not be constructive to frame the use of sensor data in terms of a variable that has been collected for the last 20 years and has a particular set of known error properties. The sensor is not going to collect the same information in exactly the same way—it may measure something a little different or have different error properties. One way forward, as a sort of hybrid, is to continue the current expensive collection on a smaller scale and use less expensive but more error-prone sensor data on the majority of the sample. This may provide a basis for modeling what the dataset would have looked like if the more expensive method had been used for everybody.

Brown then turned the conversation to questions related to transparency. Specifically, the question to the audience was how they (or their organizations) were transparent around work that is not going well. Is this sort of communication something that should be prioritized? What can be done to help promote communication across data producers in the future?

The response from the audience centered on the need for a safe space to be able to share negative information collaboratively. One member of the audience commented that maybe when an agency reports things that aren't going well, it helps to give legitimacy to the things that are reported to be going well. Much of the discussion, however, focused on the cost of making a mistake. Some errors may be more costly than others. This is something that the agencies need to understand so that resources can be focused appropriately. Additionally, the question of how many rabbit holes the agencies want to share with the end user was raised. There is a cost to producing the material to explain an error and a cost to the end user reading and absorbing that material.

To wrap-up, Brown asked the general question of how agencies assess the value to users of the data they produce. At the National Agricultural Statistical Service (NASS), feedback is solicited through an annual advisory committee meeting as well as a data user conference at which NASS staff meet directly with data users. Neither are formal, but both channels provide substantial amounts of information that inform how the agency proceeds. The final comment was that much of the literature on standards distinguishes among different types of users – what needs to be reported to the least sophisticated user may be quite different from what needs to be reported to the most sophisticated user. The challenge is that the agencies may not be able to communicate enough to the most sophisticated users and thus lose the feedback that those users could provide.

References

Abowd, J., Stinson, M. (2011). Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Bureau Survey and SSA Administrative Data. SEHSD Working Paper (2011-19).

Bradley J, Wikle C, Holan S. (2016). Bayesian Spatial Change of Support for Count-Valued Survey Data. *JASA*

Bradley, J, Wikle, C, Holan S. (2015) Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates. Stat

Bradley, Wikle and Holan (2017). Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error. JRSS-B

Daas, Piet, and Saskia Ossen. (2011) *Deliverable 4.2: Report on Methods Preferred for the Quality Indicators of Administrative Data Sources,* Report for Work Package 4 of the European Commission 7th Framework Program BLUE-ETS. Brussels, Belgium: European Commission,.

Dinur, I., Nissim, K. (2003). Revealing Information while Preserving Privacy. PODS 2003, June 9-12, , San Diego, CA

European Commission. (2011). *European Statistics Code of Practice*. Brussels, Belgium: European Statistical System Committee,

European Statistical System Committee.(2015) *Quality Assurance Framework of the European Statistical System, Version 1.2.*.http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646

Eurostat. *ESS Handbook for Quality Reports*, 2014 edition. Luxembourg: Publications Office of the European Union, 2015.

Fosdick, Bailey K.; DeYoreo, Maria; Reiter, Jerome P. (2016), Categorical data fusion using auxiliary information. Ann. Appl. Stat. 10no. 4, 1907--1929. doi:10.1214/16-AOAS925. https://projecteuclid.org/euclid.aoas/1483606845

Groves, Robert M., F. J. Fowler, Jr., Mick Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourrangeau.*Survey Methodology*. *Revised edition.* New York: Wiley, 2009.

Groves, Robert M, and Brian A. Harris-Kojetin. Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy. 2017. Print.

Holmberg, Anders, and Christine Bycroft. “Statistics New Zealand’s Approach to Making Use of Alternative Data Sources in a New Era of Integrated Data.” In Paul P. Biemer et al., editors, *Total Survey Error in Practice*. New York: John Wiley and Sons, 2017.

Kasprzyk, Daniel, and Graham Kalton. “Quality Profiles in U.S. Statistical Agencies.” *Proceedings of the International Conference on Quality in Official Statistics,* Stockholm, Sweden, 2001.

Lohr, S. L. and Brick, J. M. (2012). Blending domain estimates from two victimization surveys with possible bias. *Canad*. *J*. *Statist*. 40 679–696.

Lohr, Sharon L.; Raghunathan, Trivellore E. (2017) Combining Survey Data with Other Data Sources. Statist. Sci. 32, no. 2, 293--312. doi:10.1214/16-STS584. https://projecteuclid.org/euclid.ss/1494489817

Mary K. Muth, Megan Sweitzer, Derick Brown, Kristen Capogrossi, Shawn Karns, David Levin, Abigail Okrent, Peter Siegel, and Chen Zhen. Understanding IRI Household Based and Store-Based Scanner Data, TB-1942, U.S. Department of Agriculture, Economic Research Service, April 2016.https://www.ers.usda.gov/publications/pub-details/?pubid=47636

Murff HJ, FitzHenry F, Matheny ME, Gentry, N, Kolter, KK, Crimin, K, Dittus, R, Rosen, A, Elkin, P, Brown, S and Speroff, T. (2011) Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing. *JAMA.* 2011;306 (8):848–855. doi:10.1001/jama.2011.1204

National Academies of Sciences, Engineering, and Medicine. (2017) *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24893.

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. doi: 10.17226/24652.

Oberski, D., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*.

Reid, Giles, Felipa Zabala, and Anders Holmberg. (2017) “Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ.” *Journal of Official Statistics*, vol. 33, no. 2 , pp. 477-511.

Statistics New Zealand. (2016).*Guide to Reporting on Administrative Data Quality.* Wellington, NZ: Statistics New Zealand

Steorts, Ventura, Sadinle, and Fienberg (2014). "A Comparison of Blocking Methods for Record Linkage." *Privacy in Statistical Databases (Lecture Notes in Computer Science 8744)*, ed. J. Domingo-Ferrer, Springer, 253-268.

Sweeney L, Von Lownfeldt M, and Perry M. (2017) Saying it’s Anonymous Doesn’t Make It So: re-identifications of “anonymized” law school data. Harvard University. Data Privacy Lab. White paper.

Sweeney. (2000) Simple Demographics Often Identify People Uniquely. dataprivacylab.org/projects/identifiability/index.html

Sweeney. (2015). Only You, Your Doctor, and Many Others May Know, Technology Science. 09/29/2015. https://techscience.org/a/2015092903/

Sweeney L, Crosas M, Bar-Sinai, M. (2015) Sharing Sensitive Data with Confidence: The datatags System. Technology Science.

Sweeney L, Yoo J, Perovich L, Boronow K, Brown P, Brody J. (2017) Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Technology Science.

US Commission on Evidence-Based Policy Making. (2017) CEP Final Report: The Promise of Evidence-Based Policymaking. https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

Yoo, J, Thaler A, Sweeney, L and Zang, J. (2017) Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science.

Sweitzer, Megan, Derick Brown, Shawn Karns, Mary K. Muth, Peter Siegel, Chen Zhen. Food-at-Home Expenditures: Comparing Commercial Household Scanner Data From IRI and Government Survey Data, TB-1946, U.S. Department of Agriculture, Economic Research Service, September 2017.https://www.ers.usda.gov/publications/pub-details/?pubid=85251

United Nations Economic Commission for Europe. *A Suggested Framework for the Quality of Big Data*. New York and Geneva: United Nations, 2014.

William Iwig, Michael Berning, Paul Marck and Mark Prell. Data Quality Assessment Tool for Administrative Data. Report produced by Federal Committee on Statistical Methodology (FCSM). February 2013. Accessed on 12/28/17 https://stats.bls.gov/osmr/datatool.pdf

Zhang, Li-Chun. “Topics of statistical theory for register-based statistics and data integration.” *Statistica Neerlandica*, vol. 66, no. 1 (2012), pp. 41-63.