Lightning Talk and Poster Abstracts

Lightning Talk Abstracts:

Andrea Carlson- *USDA, Economic Research Service*

Elina Tselepidakis Page, Kevin Kucynski, TusaRebecca Pannucci, Kristen Koegel, Thea Palmer Summerman, Carine E. Tornow, Sigurd Hermansen

Importing nutrients into scanner data and estimating prices for foods in the National Health and Nutrition Examination Survey: The Purchase to Plate Crosswalk and Food Price Tool

Americans spend about half of their food budgets to purchase about two-thirds of their food from stores. While many factors such as taste, level of convenience, and healthfulness impact food decisions, price is also an important factor. To conduct research on food choices, USDA purchases retail and household scanner data and maintains extensive nutrition and dietary intake databases including What We Eat in America, the dietary component of the National Health and Nutrition Examination Survey (WWEIA/NHANES). These data also support federal food programs and regulations, such as estimation of the Thrifty Food Plan, which sets the maximum allotment for the Supplemental Nutrition Assistance Program (SNAP).

Scanner data contain detailed product and purchase information, but the nutrition data are not sufficient to measure how well Americans follow dietary advice or what may motivate them to do so. On the other hand, WWEIA/NHANES provides extensive nutrition and health information and the Flexible Consumer Behavior Survey (FCBS) collects information on NHANES participants' knowledge, attitudes, and beliefs regarding nutrition and food choices, but NHANES does not include food prices.

We solve both problems by using probabilistic and semantic matches to merge the scanner data with the USDA nutrition and dietary intake databases to create the Purchase to Plate Crosswalk (PPC), allowing scanner data users to import USDA nutrition data into the IRI scanner data. The PPC covers about 95 percent of the sales in both the retail and household data. We then create the Purchase to Plate Price Tool (PPPT) to estimates prices for foods Americans report eating in the dietary intake study WWEIA/NHANES. The PPPT calculates a price for about 97 percent of the food mentions in WWEIA. Researchers can also use the PPPT with a subset of the scanner data to produce prices appropriate for their research question.

John Voorheis - *U.S. Census Bureau*

Garret Christensen, Laura Erhard, Thesia Gerner, Brett McBride, Nikolas Pharris-Ciurej

The Promises and Challenges of Linked Rent Data from the Consumer Expenditure Survey and Housing and Urban Development

The Bureau of Labor Statistics' Consumer Expenditure Survey (CE) is the primary source of data on the basket of goods and services that is used to calculate the Consumer Price Index. Concerns about rising costs and response burden have spurred research into whether administrative records can be used to improve the CE. This paper matches administrative data from Housing and Urban Development (HUD) to CE data to investigate alignment between survey responses to questions about rental assistance receipt and monthly rent paid with administrative records data. We first link all CE sample units to the HUD data, and find that sampled HUD participants are more likely to respond to the CE than are non-HUD participants. Then, we match CE respondents to HUD administrative data, which results in match rates which line up with the overall rates of public housing utilization in the US, and a matched sample with relatively more non-white householders and female-headed households. When we compare the amount of monthly rent reported in the CE compared to HUD administrative records, we find an interesting pattern: the distribution of the difference between HUD reported total tenant payment (the amount less any subsidy) and CE reported rent is skewed toward negative values, suggesting that CE rent may be over-reported for the linked sample. We investigate how this misreporting may affect estimates produced from survey data. We find that calculations of the proportion of individuals experiencing rent burden are substantially smaller when using linked CE-HUD data than when using CE survey data alone. Additionally, in a work in progress, we investigate how estimation of Supplemental Poverty Measure poverty thresholds would be affected by replacing survey-reported rent with HUD administrative data on gross rent.

Emily Wiegand - *Chapin Hall at The University of Chicago*

Robert Goerge, Nicole Deterding

Complexities and Practical Solutions for Linking Human Services Datasets

There is a great deal of interest in the public sector and among researchers in linking public sector administrative datasets from state and local agencies to better understand the impacts of human service programs, target services, and assess questions of family wellbeing. While the methodology used to link datasets has implications for analyses conducted with the linked data, there are no documented best practices for linking human services datasets, and most record linkage research is concentrated on medical data, population-level data sources such as vital records, or private sector use cases. This project explores and defines a "human services use case" for record linkage and develops an argument for discussion between practitioners, researchers, and data providers about quality and rigor in human services record linkage.

The project, part of the HHS-funded Family Self-Sufficiency Data Center, combined a series of interviews with record linkage practitioners and data experts routinely working with state and local administrative data in the human services with a literature review of available methodologies and a limited survey of commercial products.
Characteristics of this use case include very limited identifiers and unknown rates of overlap between datasets. It is also common for the quality of a given data element to vary across records in the same dataset. Each of these represents a significant risk to the quality of the

resulting link, while making it simultaneously more challenging to assess the quality of a given match.

Currently, subject matter experts create high quality matches in the human services space through very manual processes that make use of deep content knowledge. New record linkage methods such as supervised classifiers and collective matching techniques may provide opportunities to incorporate additional richness from the datasets in a more scalable fashion, but will come at an increased cost of transparency.

Suad El Burai Félix - *Centers for Disease Control and Prevention / National Center for Health Statistics*

Cordell Golden, Cindy Zhang, Christine Cox, Lisa B. Mirel

Comparative Analysis of the NHANES Public-Use and Restricted-Use Linked Mortality Files

Linking national survey data with administrative data sources enables researchers to conduct analyses that would not be possible with each data source alone. Recently the Data Linkage Program at the National Center for Health Statistics (NCHS) released updated linked mortality files, including the Continuous National Health and Nutrition Examination Survey data linked to the National Death Index mortality files. Two versions of the files were released, a restricted-use file available through the NCHS Research Data Center and a public-use file. To reduce the re-identification risk, statistical disclosure limitation methods were applied to the public-use files before they were released. This included limiting the amount of mortality information available and perturbing cause of death and follow-up time for select records. To demonstrate the comparability between the restricted and public-use files, relative hazard ratios for all-cause and cause-specific mortality, using Cox proportional hazards models, were compared and the absolute relative bias was assessed. This presentation will describe the importance of such work in the context of data quality and fitness for use.

Mark Motivans - *Bureau of Justice Statistics*

Using Linked Data from the Federal Justice Statistics Program

The Federal Justice Statistics Program is managed by the Bureau of Justice Statistics (BJS) and serves as the national clearinghouse of administrative federal criminal case processing data. Under this program, data are received from six federal justice agencies each year and are standardized, maintained, linked, analyzed, and archived. BJS uses identifiers from these data sources to create link files that offer researchers interested in studying the case processing of federal defendants a new way to investigate a range of issues affecting policy, theory, and practice. This poster describes the linked files—a set of cross-walks that provide the researcher with the ability to track suspects forward (e.g., from arrest to subsequent processing stages) and backward (e.g., from sentencing to earlier stages); the link rates, and potential applications for enhancing federal statistics. The goal is to promote the use of the BJS FJSP data and its linking

system for external researchers and to share our experiences with others who have similar issues linking records across diverse data.

Amy Burke - *National Science Foundation*

Leigh Ann Pennington, Mike Finn

Estimating Stay Rates of Foreign-Born Science and Engineering Doctorate Holders by Using Data from the National Center for Science and Engineering Statistics and the Social Security Administration: 1995 to Present Day

Foreign nationals have received about 40% of all science and engineering (S&E) doctorates earned in the United States since 2000.Via a collaborative agreement with the Oak Ridge Institute for Science and Education (ORISE), the National Center for Science and Engineering Statistics (NCSES) has been collecting information on whether these graduates remain in the United States for 5, 10, and 16 years after receiving their degrees. Since 1995, NCSES and ORISE have linked data from the NCSES Survey of Earned Doctorates (SED) with Social Security Administration (SSA) data to calculate "stay rates" over time for foreign-born S&E doctorate holders with temporary visas. This linking method was first used for S&E doctorate-holders graduating in years 1984 through 1988 and who reported earnings in 1992. This method has been used to calculate updated stay rates every 2 years up through those graduating in 2006 with earnings as recent as 2015. Use of aggregated data based on characteristics of interest, such as degree field and country of citizenship, rather than individual-level records, eliminate most privacy concerns. The method has been validated with data on immediate post-degree plans collected in the SED, and recently through calculations of the stay rates via data from the NCSES Survey of Doctorate Recipients, whose population was adjusted in 2011 to allow for calculation of a stay rate. Recent challenges include matching the SSA data based on only the last-four digits of the social security number, as now collected in the NCSES SED.

Daniel Flynn - *U.S. Department of Transportation*

Michelle M. Gilmore, Erika A, Sudderth, Pat Hu, Ed Strocko, Paul Teicher

Linking crowdsourced and traditional traffic incident data to estimate roadway crash counts

Crowdsourced mobile applications such as Waze are an emerging and rich source of real-time data that reflect roadway conditions, when and where users are active. We assessed the potential for Waze alerts to serve as a reliable and timely indicator of police-reportable crashes, which may support short-term monitoring of safety performance. We integrated Waze data with a set of relevant safety and contextual data to estimate police-reportable traffic crashes within one-mile-area grid cells across four states, using R, Python, and ArcGIS. We linked hourly Waze alert data with roadway geometry, traffic volume, vehicle crashes, demographic, economic, and daily weather data over a one-year period for each grid cell. We found that including Waze alerts significantly improved the performance of the state-wide crash estimation models. The specific spatial and temporal patterns of the estimated crashes from the models is

close to the observed police reported crashes, but not exact. For example, the models underestimate crashes during early morning hours by an average of 2-10% across the four states, and overestimates crashes by 0.8-8% at commuting times, when the volume of Waze data is highest. The statewide Waze crash models appear to capture unreported crashes, including minor crashes which might not generate a police report, but can seriously impact congestion. To test specific applications, we applied the Waze data integration and modeling methods in two case studies. In collaboration with the Tennessee Highway Patrol, we found that integrating Waze data significantly improved the spatial and temporal resolution of the existing crash propensity models used to prioritize patrol locations. We also worked with the city of Bellevue, WA to integrate Waze alerts, 911 traffic incident records, and historical crash data with segment-level roadway features. We developed interactive road network dashboards to support Bellevue's Vision Zero action plan. The case studies offer specific examples of how crowdsourced traffic data such as Waze can enhance other roadway data to illuminate safety risk patterns and inform decision making.

Emanuel Bendavid - *U.S. Census Bureau*

Martin Slawski

Regression with linked data files

The Census Bureau regularly combines data from a variety of sources such as administrative data, survey and census data. The main purpose of combining data is to reuse existing data, reduce the cost of data collection, research and burden on responders. In the absence of unique and common identifiers across such data files, record linkage (or data matching) is an essential task to identify which records in different data files belong to the same entity. Because record linkage is prone to linkage error- false matches and false non-matches- statistical analysis of linked data files can suffer from selection bias and adverse outliers. Therefore, to adjust the statistical analysis, it is of interest to develop statistical methods that can alleviate the adverse effects of linkage error. In this talk, I consider the regression analysis of linked data files and propose several approaches for estimating the regression parameters and also identifying and correcting falsely matched records.

Poster Session Abstracts:

Scott Wentland - *Bureau of Economic Analysis*

Zachary H. Ancona, Kenneth J. Bagstad, James Boyd, Julie L. Hass, Marina Gindelsky, Jeremy G. Moulton

Accounting for Land in the United States: Integrating Physical Land Cover, Land Use, and Monetary Valuation

The United States does not yet produce formal accounts to quantify the aggregate value of its privately owned land. We develop a pilot set of national and sub-national land accounts, which include an initial set of both physical and monetary accounting of land in the U.S. Methodologically, we show that it is feasible to produce land values that can be directly linked to and integrated with physical land cover/use data. The physical accounts utilize detailed land use (National Land Use Database) and land cover (National Land Cover Database) datasets, which provide insights into how land cover and use in the U.S. are changing over time. To provide aggregate estimates of land values, we use a hedonic regression approach that exploits fine-grain microdata ("big data" from Zillow) that contains detailed information from hundreds of millions of property transactions and their corresponding physical characteristics covering much of the U.S. Indeed, we show that all three data sources can be used to generate aggregate estimates of land value by directly incorporating linked land-cover data into valuation models that use the Zillow microdata. We then use categorical land-use data as physical quantities for constructing regional and national estimates. The extent to which we "cross-pollinate" separately collected physical and monetary "big data" to construct national land-value estimates from the bottom-up is itself a novel addition to the literature and could be adapted for use in other countries as similar "big data" becomes increasingly available to national statistical offices. Overall, these pilot accounts reveal that U.S. land cover has shown declines in forests, cropland, and pasture with increases in barren, scrub/shrub, and developed classes, which are particularly concentrated in the U.S. Southeast. We also find that nominal land values in the U.S. fell about 30% from the boom to bust periods in the prior decade, albeit with substantial regional variation, and have subsequently experienced a nearly full recovery in recent years.

Jonathan Rothbaum - *U.S. Census Bureau*

Charles Adam Bee

Administrative Income Statistics Project: Using Linked Data to Improve Income Statistics

There has been considerable interest within the U.S. Census Bureau and in the research and policy communities regarding error in the measurement of income on surveys. One possible avenue to correct for this error is to use administrative data to replace or complement survey responses. We summarize recent research conducted at the U.S. Census Bureau on how survey estimates of various income statistics may be biased by issues such as misreporting and increasing non-response. For example, Bee and Mitchell (2019) find that for those 65 and older, median household income is understated by about 30 percent and poverty is overstated by

about 25 percent due to underreporting, primarily of retirement income. Hokayem, Raghunathan, and Rothbaum (2019) find that non-response to income questions causes official estimates of overall median household income to be overstated by 1 to 5 percent and estimates of poverty to be understated by 0.5 percentage points. Other recent work has evaluated unit non-response bias by linking addresses in the survey frame to administrative records (Bee, Gathright, and Meyer, 2015). We detail the implementation opportunities and challenges to using administrative data. Finally, we outline our research agenda to address these challenges.

Xiaoshu Zhu - *Westat*

David Morganstein, Sarah Shore, Frost Hubbard

An application of the record linkage macro WesLink to identify duplicates within and across multiple sampling frames

The Commercial Buildings Energy Consumption Survey (CBECS) uses multiple sampling frames to collect energy-related building characteristics and energy usage data. We have to create some of these frames while others are acquired. As a result, the same building may appear within and across frames, and therefore, has multiple chances to be selected. To control the overlap and compute the selection probabilities of buildings, we apply a self-developed record linkage program called WesLink to identify duplicate records. The WesLink is a SAS macro which uses probabilistic methods to identify records that belong to the same entity. In this study, we describe the process of de-duplication and matching under an established hierarchy of lists developed in previous rounds of CBECS, with special focus on the features of WesLink, including selecting matching fields, establishing parameters for matching weights, determining threshold for matching pairs and evaluating link quality. We compare the matching rates of WesLink to the manual review results from the 2012 CBECS and discuss the advantage of using WesLink to improve matching efficiency while maintaining the same level of accuracy.

Ricardo Limes - *U.S. Bureau of Economic Analysis*

Ryan Noonan

BEA and BLS Blend Data to Gain New Insights about Foreign Direct Investment in the United States

Many data users are curious about the effects of globalization on the U.S. economy. A recent collaboration between the Bureau of Economic Analysis (BEA) and the Bureau of Labor Statistics (BLS) helps shed light on the segment of the American workforce employed by foreign multinational companies. BEA combined its wealth of survey data on foreign-owned U.S. businesses with the BLS Quarterly Census of Employment and Wages (QCEW) and Occupational Employment Statistics (OES) to uncover new insights on employment, wages, and occupations for foreign-owned companies:

- Did you know that Ohio is home to the top two counties in the country in terms of employment attributed to foreign-owned companies? Foreign-owned companies employ 40 percent of workers in Union County and 34 percent in Logan County.

- Or that STEM (science, technology, engineering, and mathematics) occupations make up nearly 13 percent of employment in foreign-owned companies, compared with 6 percent in domestically-owned companies?

By blending these existing data sets, BEA and BLS produced new information at the national, state, and local areas, as well as additional industry-level detail, without increasing public burden. This case study shows the opportunities of cross-agency data collaboration, as well as some of the challenges of using big data and administrative data in the federal government.

Jennifer Ortman - *U.S. Census Bureau*

Sandra L. Clark, Nikolas Pharris-Ciurej

Big Survey Meets Big Data: Implementing Administrative Records on the American Community Survey

The U.S. Census Bureau has made significant progress exploring the use of administrative records in the American Community Survey (ACS) to continue to meet data needs in an era of limited budgets, rising costs and decreasing participation. Incorporating administrative records into our processes should positively impact respondent burden and data reliability, while saving costs by, for example, reducing the need for follow up visits. There is great potential for administrative record utilization in data collection and processing, but there are also great challenges. These include matching accuracy, geographic coverage, and a mismatch between administrative concepts and statistical requirements. This paper details the vision of how administrative data will be integrated into the ACS, including an evaluation of alternative administrative data sets, a case study on the use of administrative data to replace ACS housing items, and the use of administrative data for editing and imputation on the ACS.

Kate McNamara - *U.S. Census Bureau*

Kaitlyn King

Census Bureau Evidence Building Projects

The Evidence Building Staff at the Census Bureau facilitates the use of administrative data as evidence for program evaluation. Previously, administrative records have been incorporated and made available for researchers' use in a limited way via the Census Bureau's research data centers (RDCs). Certain legal requirements and other parameters severely reduced evaluators' abilities to conduct evaluation studies using linked administrative records in the RDCs. Recent efforts in evidence-based policymaking, including the Foundations of Evidence Based Policymaking Act of 2018, now offer new opportunities for data linkage and program evaluation. The work of the Evidence Building staff aims to improve access to the data linkage infrastructure for program evaluation in collaboration with outside researchers and other agencies. There are several projects underway at the Census Bureau focused on using data linkage to evaluate federal and other programs. Examples of current collaborative projects

include work with FEMA, HUD, VA, Chapin Hall at The University of Chicago, and several other universities.

Julia Redmon - *U.S. Department of Education*

Patrick Keaton, Beth Sinclair

Common Core of Data (CCD): Linking the public district and school universe to other data sources

The Common Core of Data (CCD) is the universe of public schools and districts in the U.S. There are many opportunities to link CCD with other education data using NCES IDs (unique identifiers) and even more untapped opportunities to link to data outside of the U.S. Department of Education. The poster presentation will include recommended standard linking process with sources that contain NCES IDs, examples of ways CCD data are used today, and opportunities and methods to link CCD and EDFacts data with other Agency data sources.

Liana Fox - *U.S. Census Bureau*

Jonathan Rothbaum

Correcting for SNAP Underreporting: Using Administrative Records to Develop a Model for Improving Reported SNAP Receipt in the CPS ASEC

Given survey mis-reporting, linked data offers the opportunity to improve estimates of income and resource statistics as well as evaluations of program impacts. However, in many cases, administrative data is not available in all locations. For example, SNAP, WIC, and TANF data are currently available for some states in some years. Even if we assume that administrative data are relatively free of error, we still must address mis-reporting in the locations for which we do not have administrative records. Any approach to adjust for survey under-reporting in the absence of administrative microdata must make assumptions about the relationship between benefit mis-reporting and household and individual characteristics. Without access to linked microdata, two well-known techniques, Urban Institute's Transfer Income Model (TRIM) and CBO's regression-based adjustment, require assumptions that are difficult to verify. In this paper, we treat the problem of incomplete geographic coverage of administrative data as a standard case of missing information. It is as if all individuals in particular locations refused to answer the question, "What do administrative records in your state SNAP office say about your participation in the program?" We then impute responses by modeling administrative program data given the observable information from states where we do have administrative data to states where we do not. The advantage of this approach over alternative approaches is that the underlying assumption can be tested. We compare estimates using the administrative data to estimates produced using the imputed data. For example, if we have program data from New York, we can treat that data as missing and impute SNAP participation and amount received using the other state data. We then compare summary statistics and regression coefficients for New York using the imputed data and the actual administrative records to test whether this imputation approach yields unbiased results.

Keith Finlay - *U.S. Census Bureau*

Mike Mueller-Smith

Criminal Justice Administrative Records System (CJARS)

The Criminal Justice Administrative Records System (CJARS) is a Census Bureau project started in 2016 to create a national, integrated, harmonized collection of criminal justice microdata at the Census Bureau. The project is a joint effort with the Population Studies Center at the University of Michigan. The project has three fundamental goals: (1) improve Census Bureau operations, (2) provide valuable aggregate statistical information to criminal justice agencies, and (3) increase the quality and quantity of criminal justice research by making the data available through the Federal Statistical Research Data Centers.

CJARS has collected data on 57 million criminal justice events from agencies in 14 states dating back to the 1970s. These records include arrests, court proceedings, and periods of probation, prison, and parole on 17 million unique individuals. PII in these data allow CJARS to connect individuals through the stream of criminal justice events.

The project is designed to be sustainable from its foundation, providing value to all participants. Data providers receive reports using linked data, including on non-criminal justice outcomes, to facilitate public administration. The federal statistical system improves its operations and data quality related to the criminal justice system. Researchers gain access to a uniform data infrastructure, and their research will ultimately better inform policymakers about how the criminal justice system functions.

Many of the most interesting applications of CJARS data are the opportunities for data linkage with non-criminal justice data. Links with the SSA Numident file allow us to identify the cumulative likelihood of criminal justice involvement given an individual's place of birth. Links with IRS tax filings will facilitate research on the interactions between labor force participation, crime, and sanctions. Survey and administrative data held at Census will support research on the impacts of criminal justice involvement on families.

Matthew Bouch - *U.S. Census Bureau*

Data Linkage with SK-Learn

The purpose of this project was to perform a brief evaluation of SK-Learn methods for data linkage purposes. The project used publicly available record linkage data from the UCI Machine Learning Repository. SK-Learn has many methods which can be used for machine learning purposes. However, the methods that were chosen for this project were LogisticRegression, SVC, DecisionTreeClassifier, MLPClassifier, and RandForestClassifier. Each of these methods were used to produce models that were tuned, fitted and evaluated. The models were subjected to only trivial hyper-parameter tuning, using GridSearchCV to discover slight adjustments to the defaults. The chosen models were all able to perform at similar and acceptable levels. Each model achieved an ROC-AUC score exceeding 0.99. An extension of

the project was to create data linkage software which used the stored SK-Learn models. The Pandas library was used to read data, group data by user specified blocking factors, create feature arrays, and then apply the models to the candidate pairs. This approach makes performing data linkage, using SK-Learn and Pandas, simple to perform.

Karen White - *National Science Foundation*

Wan-Ying Chang, Cassidy Sugimoto

Demographic Differences in the Publication Output of U.S. Doctorate Recipients

We link administrative records on peer-reviewed publications (Web of Science (WoS)) and national survey data (Survey of Doctorate Recipients (SDR)) to investigate whether demographic differences among U.S. Doctorate Recipients impact publication output. The project employed an innovative method for linking databases and has revealed insights into author demographics which were previously not available.

Innovative Data Linking. The SDR respondents are matched to the authors of publications indexed by the WoS using a machine learning approach. The data linkage is challenging because while the SDR names are disambiguated as to names, the WoS is not. We constructed a gold set of matches with high confidence; used Random ForestTM model to identify publications that could be matched; created seed publications that were matched to SDR respondents; and refined the data to ensure no more than one authorship on a single publication, and classified email matches as extract matches.

Insights from Analysis. Our research shows the most impactful determinants on the probability to publish are related to field of doctorate, employment sector and engagement in R&D activity. A doctorate recipient's training is also significant, those who complete dissertations at high-research institutions are more likely to publish. After controlling for the dominant determinants, the demographic variables, race/ethnicity, gender, and U.S. citizenship status at the time of graduation, still show impact on the probability of publishing. The ability to test a broad range of demographic variables is unique to the SDR-WoS dataset. Of the demographic variables, race/ethnicity has the strongest impact on likelihood to publish.

Gail Mulligan - *U.S. Department of Education*

Elise Christopher, Carolyn Fidelman, Tracy Hunt-White, Jill McCarroll

Developing from within: Using NCES Administrative Data to Build NCES School Sampling Frames and Study Data Files

The sample surveys within the Longitudinal Surveys Branch at the National Center for Education Statistics (NCES) utilize administrative data from NCES's universe data collections of K-12 and postsecondary education institutions in various ways. The Common Core of Data (CCD) is a universe data collection of all K-12 public schools in the United States, while the Private School Universe Survey (PSS) is a universe data collection of all K-12 private schools. The

elementary, middle, and secondary longitudinal studies use school information from these two collections to build a sampling frame from which to sample schools. The National Postsecondary Student Aid Study (NPSAS) develops its sampling frame from the Integrated Postsecondary Education Data System (IPEDS), a universe data collection of postsecondary institutions. In addition to using the universe files to develop sampling frames, the longitudinal studies include data about school characteristics collected in these universe collections in their own study data files that are disseminated to researchers. This poster will discuss the process of developing a sampling frame using these data files, including information used to sample by strata and procedures for freshening to account for newly opened schools. It will also discuss the types of school information included in the data files released to researchers, including the advantages and disadvantages of relying on the administrative data instead of collecting information directly from schools during each longitudinal study's data collection.

* As a note about whether this work is published, most of the details to be presented are included in various locations in study documentation but not in a focused way with the emphasis on linkages as would be presented in this poster.

David Popko - *Bureau of Labor Statistics*

Ilmo Sung

Does location matter? A case-study of the influence of geography in measurement of gasoline price inflation

The Consumer Price Index (CPI) objective is to measure the change in cost of living experienced by the average urban consumer residing in the United States. Currently, the outlet sample in this measurement is selected via probability sampling proportional to average daily expenditure on items within a core-based-statistical-area (CBSA). This process yields a large variety of outlets in the CPI sample, but the outlets selected remain exclusively a function of the expenditures reported and household sampling weights in CPI's Telephone-Point-of-Purchase Survey (TPOPS) survey.

Using data collected from GasBuddy.com, we attempt to model the explanatory variables of price change in order to identify possible stratification variables in outlet selection. Focusing on the Washington-Arlington-Alexandria, DC-VA-MD-WV PSU, we calculate one-month price changes for each gas station in the GasBuddy sample. We then apply various statistical techniques to assess the significance of a variety of independent variables constructed using traffic data, driving distances between, population density, income, housing prices, and other while controlling for various geographic flags. Finally, we construct indexes from the GasBuddy sample using a proposed stratified methodology, and compare them with indexes constructed using the traditional CPI methodology.

Kenneth Jones - *Department of Veterans Affairs*

Ernest M. Moy

Equity-Guided Improvement Strategy to Reduce and Eliminate Veteran Health Disparities

The Veterans Health Administration (VHA) Office of Health Equity (OHE) improves the health and wellbeing of Veterans by advancing the reduction and elimination of health disparities and inequities. This entails incorporating and promoting the collection of data and developing tools to better understand where disparities and inequities may exist for Veterans receiving care in VHA facilities and in the community. OHE has launched an evidence-based and data strategy as a set of consultative tools to assist Veterans, medical providers, and stakeholders to make informed care decisions and inform local and national strategies to reduce, and ultimately eliminate, disparities when detected in Veteran populations. This strategy aligns with the VA MISSION Act that provides the Department of Veteran Affairs (VA) the tools to provide care to Veterans that is trusted, easily accessible, and high quality. The purpose of the presentation is to showcase a set of tools, including data stories and visualizations, that use linked data to highlight Veteran care and health equity issues at VHA facilities and in places where Veterans live, work, and play regardless of whether these Veterans receive VA care or benefits. Highlighted data tools include the use of clinical quality measures and other administrative data to produce a series of national health equity reports and data visualizations, annual Veteran projection models to highlight changes in care needs for vulnerable Veteran groups, combining data from projection models to show potential impacts of social and economic determinants of health at various geographies, and using data from VHA facilities and benchmarking data from stakeholders to spotlight facilities recognized for their commitment to equity and inclusion.

Edward Porter *- U.S. Census Bureau*

False Duplicates in the Census: A novel approach to identifying false matches from Record Linkage Software

Identification of duplicate records of individuals is necessary for an accurate census. The removal of duplicates in the 2010 US Census involved using record linkage software. That software compared two responses from the census and assigned each pair a matching weight. The higher weights would indicate the likelihood of a duplicate. Using software, which grouped the individual records by the response, members of the decennial staff review the census responses and manually set one cutoff weight. Pairs of responses, with a greater matching weight than the cutoff, were classified as duplicates. All other responses that were not classified as duplicates would then be linked or not linked as individual records. This poster describes a novel approach to identify those responses that were incorrectly assigned as matches. This strategy is a graphical network to eliminate false matches. By creating a "social" network using responses as nodes and a classified duplicate link between the responses as edges, the largest subgraphs (cliques) could be indications of false matches. One would expect that most of the false matches would be large graphs of common surnames. However, some interesting outliers are those cliques where the names are not common. The initial work has had some unexpected knowledge discovery. These results can be used as unbiased training data for a machine learning classifier.

Paul Bernhard - *Department of Veterans Affairs - Veterans Health Administration*

Dr Yasmin Cypel, Dr JOel Culpepper, Dr Aaron Schneiderman

Geocoding Veteran Addresses to Get Distance to Nearest VA Health Facility

The 2018 Comparative Health Assessment Interview (CHAI) research study will help VA understand the effects of military service, including deployment and combat, on the health of Veterans who served during Operation Enduring Freedom, Operation Iraqi Freedom, and Operation New Dawn. The Epidemiology Program's research team in Post Deployment Health Services, Office of Patient Care Services, as well as other VA researchers, are currently investigating risk and resilience factors for mental and physical health outcomes. The CHAI dataset contains 15,166 veteran respondents (33% of whom are VA healthcare users) with self-reported mailing/residential addresses. In modelling health outcomes among different veteran groups, the need arose to control for a veteran's access to health care services. Thus, proxy geographical metrics for health care access were linked in Veterans' addresses and those of VA health facilities (n=1,513) were geocoded using the R statistical software package, "ggmap" (queries Google Maps API). Pairs of latitude-longitude were obtained. Distance from a veteran's address to the nearest VA facility was calculated using the haversine formula on a Cartesian product of the datasets. Among the results, we determined that CHAI veterans reside an average of 9.97 miles from the nearest VA facility (highest state average was 16.71 miles for South Dakota; lowest was 1.83 for DC).

Other geographical data was matched into the CHAI dataset at the ZIP or County level:

- Urban-Rural Classification Scheme for Counties [from CDC National Center for Health Statistics] (https://www.cdc.gov/nchs/data_access/urban_rural.htm)
- Rural-Urban Commuting Areas [from Department of Agriculture and the Department of Health and Human Services (DHHS)] (https://www.ruralhealth.va.gov/rural-definition.asp)
- Primary Care and Mental Health Care Shortage Areas [from Health Resources and Services Administration under DHHS] (https://data.hrsa.gov/)
- Metropolitan Statistical Area Indicator [from Census website "American FactFinder"] (https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk)

High correlation among these geographical metrics provided evidence of accuracy in the data linkage methods.

Kendra Asher - *Bureau of Labor Statistics*

Peter Meyer, Jerin Varghese

Improving Census to NAICS industry matches

The Bureau of Labor Statistics' (BLS) Productivity Program uses Current Population Survey (CPS) microdata to supplement estimates of employment and hours worked, and to adjust hours data for changes in worker education and experience. The CPS uses the Census industry classification system, whereas BLS publishes estimates in the NAICS industry classification system.

When recoding Census industry codes to three-digit NAICS codes, two key issues arise: (1) assigning observations from "not specified" Census industries and (2) assigning observations from Census industry codes that map into multiple three-digit NAICS industries. Multiple methods exist for allocating these observations.

For productivity measurement, BLS uses the Hamilton method to apportion discrete numbers among a group of "not specified" observations of self-employed and unpaid family workers. This method ensures that more source observations are recoded to NAICS industries with larger employment shares, while also controlling to aggregate employment and hours measures. In contrast, for labor composition indexes that track worker education and experience by industry, BLS removes all "not specified" observations from its estimates.

Our innovation is to use machine learning to develop algorithms to recode historical data and handle future classification changes. We explore a new method of linking that addresses issues (1) and (2) above using random forest algorithms trained on other observations. Our algorithms are based mainly on demographic and geographic factors, and in preliminary tests achieve an above 90% accuracy rate in predicting observations' three-digit NAICS industry.

Angela Wyse - *U.S. Census Bureau*

Bruce Meyer, Alexa Grunwaldt, Carla Medalia, Derek Wu

Learning about Homelessness in the U.S. Using Linked Survey and Administrative Data

Official poverty statistics and even the extreme poverty literature largely ignore the homeless. In this paper, we examine the labor market attachment, earnings, safety net utilization, demographic characteristics, and mobility patterns of individuals experiencing homelessness. This project is part of the development of the Comprehensive Income Dataset, which combines household survey data with administrative records to improve estimates of income for individuals, families and households. To allow these analyses, we link the 2010 Decennial Census, which enumerates both the sheltered and unsheltered homeless, to the 2006-2016 American Community Survey (ACS) which surveys the sheltered homeless, and longitudinal shelter use data from several major U.S. cities. We also link these data to longitudinal administrative data on the Supplemental Nutrition Assistance Program (SNAP), Temporary Assistance to Needy Families (TANF), Medicare, Medicaid, housing assistance, and mortality. We document the patterns of transitions between homeless situations and other housing statuses, as well as factors associated with these transitions. Our approach benefits from large samples that are a guide to national homelessness patterns. By shedding light on issues of data linkage and survey coverage among the homeless, this paper contributes to efforts to better incorporate this hard-to-survey population into income and poverty estimates.

Elina Page - *U.S. Department of Agriculture*

Mark Denbaly, Linda Kantor, John Kirlin

Leveraging Extant Data in the National Household Food Acquisition and Purchase Survey (FoodAPS)

The National Household Food Acquisition and Purchase Survey (FoodAPS), co-sponsored by the Economic Research Service (ERS) and Food and Nutrition Service (FNS) of the U.S. Department of Agriculture (USDA), was the first nationally representative survey of American households that integrated different types of data and information from multiple sources to provide comprehensive data on household food purchases and acquisitions. FoodAPS collected information on all foods purchased or otherwise acquired by all household members during a seven-day period, including quantities and expenditures. The survey also collected extensive information about household non-food expenditures, food security status, food assistance program participation, and access to food retailers. In addition to collecting comprehensive food acquisition information from nearly 5,000 households, FoodAPS relied on administrative data to efficiently sample hard-to-reach populations, and it linked the survey responses to extant data sources to reduce respondent burden while broadening and enriching the content. The proposed presentation will highlight the unique features of FoodAPS data with a particular emphasis on the use of extant data to enhance the FoodAPS survey data. These extant data resources include: food item descriptions from proprietary scanner data; food assistance program participation data from administrative records; nutrition data from USDA nutrient databases; store and retailer location data; store characteristics data; and food environment and food access data. The linked FoodAPS survey data have already paved the way for many new contributions to the food demand and food policy literature, particularly with regards to the diet quality of U.S. households, the impact of food assistance programs on household food choices, and the significance of the local food environment in shaping food behavior.

Zaria Tatalovich-Wenzel - *National Cancer Institute*

Benmel Liu, Donna Rivera, Mandi Yu

Linkage Initiatives in the Surveillance Research Program of the National Cancer Institute

The Surveillance Research Program (SRP) within the NCI supports the SEER Program to collect cancer data from population-based cancer registries. SRP has been increasingly engaged in record linkage initiatives to augment cancer data and facilitate a more comprehensive assessment of cancer etiology and outcomes. This poster features the following SRP data linkage projects.

Evaluating Record Linkage Software Using Synthetic Datasets: Evaluations of record linkage software using real data have restrictions partially due to limited data accessibility. To systematically test the usability of software developed for SEER - Match*Pro, we used representative model-based synthetic datasets containing patient health identifiers mimicking

the US cancer population. The project demonstrates the value of synthetic data for testing record linkage software.

SEER-SSA Birthplace Data Linkage: SEER birthplace data have substantial missing values, prohibiting their use in health disparity studies. This project aims to augment SEER birthplace data through record linkage between the SEER and Social Security Administration database containing the state of birth for US born patients and country of birth for foreign born patients. The resulting database provides sufficient information to facilitate health disparity research.

Data Acquisitions and Linkages (DAL) Initiative: The vision of DAL is to design, facilitate and evaluate large scale heterogeneous data linkages to enhance the availability of longitudinal treatment-related data for precision cancer surveillance efforts and enable epidemiologic research to improve understanding of patient health outcomes. DAL has developed a strategic framework, methodology, and initial results of linkage collaborations.

Geospatial Data Linkage: SRP provides interactive geospatial linkage tools allowing cancer control planners and researchers to facilitate analysis of contextual determinants of cancer risk and inform cancer control efforts. SRP has also begun linking cancer data with commercial residential history data to help track cancer patients who may have moved and enable research to better understand potential exposures prior to diagnosis and access to medical services.

David Shnoffner - *Social Security Administration*

Gayle Reznik, Jessie Dalrymple, Irena Dushi, Marc Sinofsky, Anya Olsen, Mark Sarney, Glenn Springstead, Joni Lavery, Pat Purcell

MINT (Modeling Income in the Near Term)

MINT (Modeling Income in the Near Term) is a micro-simulation model developed by the Social Security Administration (SSA) to analyze the characteristics of future Social Security beneficiaries and simulate the distributional effects of proposed reforms to the Social Security program. MINT is built from innovatively linking Social Security's administrative records on benefits and earnings to the Survey of Income and Program Participation (SIPP), a detailed Census survey, as well as relying on the Health and Retirement Study (HRS), the Medical Expenditures Panel Survey (MEPS), the Panel Study of Income Dynamics (PSID), the POLISIM model, and the Survey of Consumer Finance (SCF). Using sophisticated statistical methods and techniques, the model projects 21st century retirement income, marital trends, Social Security benefits, income, and poverty. The model is calibrated on the projections from the economic, demographic, and programmatic assumptions in the 2019 OASDI Trustees Report (officially "The 2019 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds"). MINT allows researchers and policy analysts to study future retirement conditions, such as income and poverty, as well as the effects of policy changes, and to provide distributional results of these outcomes to help inform policymakers. ORES's poster will provide an overview of MINT and staff will be on hand to answer questions.

Cindy Zhang - *National Center for Health Statistics, Centers for Disease Control and Prevention*

Jessica M. Keralis, Cordell Golden, Lisa B. Mirel

Mortality Experience of the 2001-2014 National Health Interview Survey Linked Mortality Files Participants

The Data Linkage Program at the National Center for Health Statistics (NCHS) regularly links its population-based survey data to data from the National Death Index (NDI) to produce linked data files containing mortality follow-up records for NCHS survey participants. The files are made available to the scientific community and offer the opportunity to study a broad range of epidemiologic research topics including the associations between mortality and a variety of health factors and determinants. Over the past 20 years, these survey-linked mortality files have been used in over 750 published research articles.

Updated survey-linked mortality files containing mortality follow-up data through December 31, 2015 were recently released. To assess the comparability of the linked survey data to the general U.S. population, we compared the mortality experience of the 2001-2014 National Health Interview Survey (NHIS) participants based on the linked mortality files with that of the general U.S. population using information from annual U.S. life tables published by the National Vital Statistics System at NCHS. Two sets of cumulative survival rates were calculated and compared, one for the annual NHIS cohort of the study participants linked to the NDI and one for the U.S. population based on the life tables.

The NHIS linked mortality files provide a unique data source for examining the associations between demographic and health factors and subsequent mortality in a large heterogeneous sample that is representative of the civilian non-institutionalized U.S. population. Understanding the similarities and differences in the estimates from the linked data and the general U.S. population is important for researchers utilizing the linked files for research. The results of the analysis will support the strength of findings of epidemiologic research using the linked mortality files.

Merianne Spencer - *National Center for Health Statistics*

Lee Anne Flagg, Geoffrey Jackson

Opioid-Involved Emergency Department Visits, Hospitalizations, and Deaths: Analysis of Linked Data from the National Center for Health Statistics

Background – Comprehensive data on opioid-involved emergency department (ED) visits, hospitalizations, and deaths are needed to inform strategies to reduce the morbidity and mortality from misuse and overdose of opioids. Data Sources – The 2014 National Hospital Care Survey (NHCS), the 2014-2015 National Death Index (NDI), and the 2014-2015 Drugs Involved Mortality (DIM) data serve as the linked data sources. NHCS collects ED and inpatient administrative claims or electronic health records from participating sampled hospitals. NDI includes data on all deaths occurring in the United States and U.S. military overseas. DIM data

provides information on the specific drugs involved in deaths based on information provided on death certificates. Methods – Opioid-involved visits were defined as having ICD-9-CM diagnosis codes 304.00–304.02, 304.70–304.72, 305.50–305.52, 760.72, 965.00–965.02, 965.09, or 970.1, or external cause of injury codes E850.00–E850.2. Drug overdose deaths (ICD-10 underlying cause-of-death codes X40-44, X60-64, X85, or Y10-Y14) with a multiple cause-of-death code of T40.0-T40.4 or T40.6 were considered opioid-involved. Unweighted results of sample study questions to examine opioid-involved ED visits, hospitalizations, and deaths due to opioid overdoses and other causes are presented. Results -- In 2014, there were 20,962 patients with an opioid-involved hospitalization that could be linked to the NDI and 1,805 died (9%) within one-year post-discharge. Of these deaths, 341 deaths (19%) resulted from a drug overdose. Of the 341 drug overdose deaths, 243 involved an opioid (71%). The opioids most involved included heroin (44%), fentanyl (20%), oxycodone (13%), methadone (12%) and morphine (12%). These categories are not mutually exclusive because a death may involve more than one drug. Conclusion – These analyses demonstrate the utility of a newly linked dataset. While the data are not nationally representative, analyses of the linked data can lend insights into hospital services and drug overdose deaths involving opioids.

Scott Campbell - *The National Opinion Research Center at the University of Chicago*

Lisa B. Mirel, Dean Resnick

Overcoming Big Data: Linking the 2014 National Hospital Care Survey to the 2014/2015 Medicare CMS Master Beneficiary Summary File

Record linkage enables survey data to be linked to other data sources, expanding the analytic potential of both the survey and the administrative data. However, depending on the number of records being linked the processing time can be prohibitive. As part of a recent project, for the National Center for Health Statistics, to link patient records from the 2014 National Hospital Care Survey to the 2014/2015 Centers for Medicare & Medicaid Services Master Beneficiary Summary File, a new method was implemented because of their size. Using SAS Enterprise Guide, a record linkage algorithm was developed based on the Fellegi-Sunter paradigm and incorporated machine learning techniques. The algorithm followed a highly structured flow and called upon several methods to improve efficiency while maintaining the integrity of the linkage. One such method used is parallel processing built on a flexible, modular coding scheme. Additional efficiency was gained by optimizing the work flow required by the record linkage blocking scheme using a machine learning approach known as sequential coverage algorithm (SCA). Utilizing a "truth source" created from a deterministic linkage matching on exact Social Security Number (SSN) agreement as the training dataset, the SCA reduced the number of linked pairs requiring evaluation while retaining a high percentage of true positive matches. Pairs generated by the optimized workflow were then evaluated by summing agreement pattern weights which were computed as a function of agreement/non-agreement probabilities. A logistic regression model, using SSN agreement as a proxy for match validity, was used to estimate probabilities of linkage validity according to the summed agreement (pair) weights. Finally, pairs were selected as links when they meet a chosen probability cutoff threshold. Using the methods described above reduced overall processing time and produced a linked file with low Type I and II error.

Derek Wu - *U.S. Census Bureau*

Bruce Meyer, Carla Medalia

Poverty in the United States Using the Comprehensive Income Dataset

This paper provides new estimates of poverty in the United States using a groundbreaking set of linked survey and administrative data. The administrative data cover earnings and asset income from IRS tax records and transfer income for a myriad of safety net programs including Social Security, SSI, SNAP, Unemployment Insurance, Veterans' Benefits, Public Assistance, housing assistance, Medicare, Medicaid, WIC, and energy assistance. We link these data to the Current Population Survey, the source of official poverty and inequality statistics and the Survey of Income and Program Participation, the most comprehensive survey of income sources in the U.S. Linking the administrative data to the surveys is vital given that a large and rising share of benefits and other income sources is not recorded in the surveys. Using these linked data, we examine the extent to which misreporting of various survey income sources biases the reported poverty status of households. We document how our knowledge of the demographics of poverty is changed by the improved data. We also provide improved estimates of other measures of the resources of the low-income population, showing how these estimates diverge from those calculated using the survey data alone.

MoonJung Cho - *U.S. Bureau of Labor Statistics*

Justin Mcillece

Practical Diagnostic Tools for Data Linkage Methods

In the Quarterly Census of Employment and Wages (QCEW) program, quarterly establishment records are linked longitudinally. Specifically, establishments that continue operations under the same ownership from quarter to quarter are linked.

In the initial steps, establishments are linked through a unique combination of state code, Unemployment Insurance Number, and Reporting Unit Number. The linkage procedure passes through a series of steps for further comparison iteratively. These steps link the vast majority of establishments between quarterly files.

For the remaining records which are not linked in the previous steps, the BLS-developed record linkage method, Weighted Match (WM), is applied. Our goal is to develop practical diagnostic tools to evaluate the performance of WM. Using decisions from manual review as the gold standard, we examined the performance of the WM algorithm in terms of a performance curve, the area under a curve (AUC), and a confusion matrix. In addition, quick and visual displays of variable scores, important variable selections using classification trees, and model fitting and prediction were provided.

Lowell Mason - *Bureau of Labor Statistics*

Record Linkage Applied to Inter-Agency Databases

It has become increasingly common to create new statistical products by integrating existing data rather than engaging in new data collection; using existing data sources is less expensive and does not increase respondent burden. An example is the integration of the Bureau of Economic Analysis enterprise-level data on Foreign Direct Investment with establishment data from the Bureau of Labor Statistics Quarterly Census of Wages and Employment. The integration of these two data sources was discussed at last year's ICSP Big Data Day (Friesenhahn, Erik. "Linking Inter-Agency Databases."). Additionally, statistical estimates derived from the integrated data sources were subsequently published: https://www.bls.gov/fdi/.

Linking separate data sources, however, involves challenges. For instance, it is usually not possible to satisfactorily link the multiple data sources without manual intervention. This may be because common identifiers do not exist in the multiple data sources, or as in this particular case, the identifiers are very noisy. For example, after linking the two data sources using the common identifiers, the initial error rate was 87.7%. After manual review and correction, the error rate was reduced to 19.0%. However, the labor cost, was considerable: 1,510.5 hours.

To reduce linkage error and labor costs, we implement several record linkage techniques that augment the common identifiers with other features of the data sources. These features, such as business names and addresses, industrial classification, and employment levels are expected to be similar but not identical among true linkages. To gauge similarity, we use a variety of similarity measures. After indexing the two data sources, we form candidate pairs and implement supervised learning techniques, such as Support Vector Machines (SVM) and Random Forests, to classify whether the candidate pairs represent a true linkage are an incorrect linkage.

Matthew Chambers - *U.S. Department of Transportation*

The Bureau of Transportation Statistics Links Data to Measure Port Performance

The Bureau of Transportation Statistic's (BTS) Port Performance Freight Statistics Program is required by law (specifically, Section 6018 of the Fixing America's Surface Transportation Act (P.L. 114-94; Dec. 4, 2015; 129 Stat. 1312)) "to provide nationally consistent measures of performance of the Nation's largest ports, and to report annually to Congress on port capacity and throughput." More than 25 Federal agencies and offices have a role in the marine transportation system (MTS), including the Nation's ports.

Our Federal partners have clearly delineated MTS roles and responsibilities such as port safety, security, energy innovation, and infrastructure investment. Painting a complete picture of port performance requires BTS to link together a host of data from our Federal partners (e.g., U.S. Army Corps of Engineers, U.S. Census Bureau, U.S. Coast Guard, and National Oceanic and Atmospheric Administration) as well as from the ports themselves.

All too often, simple questions are not easily answered, for example: what is a port, and how do you define it? Especially when our Federal partner use differing legal definition, port boundary, and even coding systems. Additionally, there are methodological differences in data collection and reporting (e.g., does moving empty shipping containers count as throughput, what's the starting point (e.g., low tide, high tide) for measuring the height of a bridge or the depth of a channel?).

BTS has had to answer many of these questions to link diverse datasets together, which were often created for specific regulatory roles. This backend linkage in a SQL database has allowed us to create interactive Port Profile visualizations, which provide comprehensive port capacity and throughput measures. For additional information on the Program and to view these Port Profiles, please visit our webpage at https://www.bts.gov/ports (click View the Port Profiles).

Erik Scherpf - *USDA*

Brian Stacy

The Implications of Misreporting for Longitudinal Studies of SNAP

Researchers studying a variety of important economics, nutrition, and health topics use survey data containing information on SNAP participation. In order to study the dynamics of SNAP participation or recognizing possible selection bias in cross-sectional estimators, many researchers use longitudinal estimators to estimate the causal effects of SNAP. However, misreporting of SNAP participation is common in survey data sets, and bias from misreporting can be larger for longitudinal estimators. In an analysis of data combining newly compiled administrative data sets on SNAP participation from nine states and covering the years 2005-2015 with individual records from the CPS ASEC survey, we confirm findings in previous studies of substantial misreporting and find evidence that the misreporting is not done at random. Additionally, we examine bias caused by misreporting in a longitudinal estimators and find severe bias, much greater in magnitude than bias caused by misreporting in cross-sectional estimators. We find that a longitudinal conditional distribution estimator may be an attractive solution for researchers using public use survey data sets.

Pheak Lim - *National Center of Veterans Analysis and Statistics*

Dr. Hyo Park

United States Eligibility Trends and Statistics (USVETS) Database

The United States Eligibility Trends and Statistics (USVETS) is an integrated multidimensional database used to support evidence-based policymaking, planning and other analytic activities. USVETS strives to capture the universe of all Veterans, living and deceased. USVETS ingests records and information from VA program offices, the Department of Defense and third-party sources. Raw data are cleaned and validated with the help of SSA validation and by applying the Jaro-Winkler scoring algorithm. USVETS integrates utilization information from VA, military

information for DoD, and socio-economic information from third-party using SAS ETL (extract, transform, load) procedures to create USVETS datasets, consisting of the Static file (cumulative) and Fiscal Year (FY) files. The current Static dataset has roughly 38.9 million Veteran records to include both living and deceased Veterans. The FY files consist of 'living' records and records that have utilization activities during the fiscal year. We have FY files from 2005 to the present, which can be used for trend analyses. Through rigorous reviews and analyses of Business Rules (BR), USVETS creates an improved data source for VetPop, VA's actuarial estimation model of the Veteran population.

Jianzhu Li - *Westat*

Tom Krenzke

Use of Probabilistic Record-Linkage to Measure Re-Identification Risk

Identity disclosure may occur if an intruder successfully matches the respondents in the survey data to the records in administrative data or other external data sources through common variables. It is therefore a recommended practice to measure the re-identification risk in the statistical disclosure control process before disseminating the data to public. Probabilistic record-linkage can be implemented to measure the re-identification risk in a potential public use file. In the Federal Employee Viewpoint Survey, we conducted probability-based matching to estimate the likelihood of correctly matching a person in a proposed partially synthetic public use file to the record in the original data. In the National Household Food Acquisition and Purchase Study (FoodAPS), we evaluated the re-identification risk due to geographical clustering. FoodAPS is a survey with a multistage stratified design, for which the variance strata and variance unit (VSVU) are included in the public use file to facilitate variance estimation. The VSVUs could potentially be used to distinguish the records in the same geographical areas (i.e., U.S. counties). As a result, in surveys of similar designs, there is potential risk to reveal the identities of the counties by matching the county-level estimates from the survey to those in the American Community Survey for a set of characteristics available in both. A probabilistic record-linkage software was used to identify the counties that were subject to high re-identification risk.

Tracy Hunt-White - *U.S. Department of Education*

Using Administrative Data to Improve Coverage of Veterans in NCES Postsecondary Sample Surveys

Starting in 2015, the National Center for Education Statistics (NCES) and the Veterans Benefits Administration (VBA) began collaborating on how to identify veterans for oversampling in NCES postsecondary sample surveys. Veterans students are an important analytical group in NCES postsecondary studies. They represent about 4 percent of all undergraduate students. In order to increase the precision of estimates for this subpopulation and to allow for more complex analyses, NCES needed to increase the sample size of veteran students. NCES has historically identified veterans using Department of Education data and institution student records. Neither of these sources provided reliable coverage of veterans to allow for oversampling. However, matching institution student enrollment lists to VBA data allowed NCES to identify veterans prior

to sampling. In addition, once sampling was completed, NCES matched its sample to VBA data to obtain the amounts of veterans education benefits received by these students. Hence, matching to VBA not only resulted in oversampling veteran students, but improved the accuracy of estimates of veterans education benefits. This poster will present information on the problems NCES has historically encountered with identifying veterans in these sample surveys, how VBA data are used to address these problems, and the challenges that emerge when using VBA administrative data in postsecondary sample surveys.

Anne Parker - *Internal Revenue Service*

Christine Oehlert

Using Machine Learning Approaches to Improve Industry Coding in the IRS Master File

NAICS Codes comprise a business classification system based on industry production processes. In the US, these codes are used to estimate Official National Industry Statistics such as GDP, Gross Output, Employment and Compensation, and Input-Output Accounts. NAICS Codes are self-reported (since 1985) on tax forms so they are subject to error. Over the ten-year period 2007 – 2016 error rates for Sole Proprietors (1040 Sch C) averaged 22% and Corporate (1120) averaged 18% at Economic Sector level (first two digits of the code). The goal of this project is to develop effective predictive models for NAICS Codes using IRS statistical and administrative data. The project uses two parallel approaches – (1) supervised models (CART, Random Forests, Boosted Trees (XGBoost)) and (2) unsupervised models (recommender algorithms). We use Statistics of Income (SOI) data over the period 2006 – 2016 linked with administrative data from Individual Returns Transaction File (IRTF) and Business Returns Transaction File (BRTF). The SOI data is a stratified probability sample with strata based on the presence or absence of a tax form or schedule and various income factors or other measures of economic size. SOI manually validates NAICS Codes and we take these validated codes as ground truth for our model development and validation. Our two approaches use statistical and administrative data in different ways. We build supervised models using statistical data augmented with administrative data. However, we build unsupervised models with only administrative data and use statistical data to validate those models. Preliminary results show accuracy rates of 72% for both supervised and unsupervised methods.

Jin Kim - *National Center of Veterans Analysis and Statistics*

Dore Glasgow

Veteran Population Model

The VetPop model links administrative data from VA and DoD with data from the US Census Bureau's American Community Survey (ACS) to estimate the total Veteran population. Although we can attain reliable data on Veterans from the VA and DoD administrative records, information on older Veterans who had separated prior to mid-1970s is limited due to missing DoD records. We use the ACS data to compensate for this data limitation. By combining these two data sources, we can more accurately estimate the number of Veterans for all ages. The

poster will show a general figure of combining records from the two data sources and then show how the results are impacted by age, gender, and period of service.