NOVEMBER 2018

ADRF Network Research Conference

INNOVATIONS *in* ADMINISTRATIVE DATA *for* SOCIAL SCIENCE

INNOVATIONS *in* ADMINISTRATIVE DATA *for* SOCIAL SCIENCE

Tuesday, November 13, 2018
9:00am – 6:00pm

Wednesday, November 14, 2018
8:30am – 4:00pm

JW Marriott Washington, DC
1331 Pennsylvania Ave NW
Washington, DC 20004

TABLE *of* CONTENTS

PROGRAM COMMITTEE

Ann Carson, Bureau of Justice Statistics
Tanvi Desai, Consultant in Data Policy and Strategy
Rashida Dorsey, Equal Employment Opportunity Commission
Dylan Ellis, Camden City School District
Kitty Smith Evans, American Economic Association
Daniel Goroff, Alfred P. Sloan Foundation
Amy Hawn Nelson, Actionable Intelligence for Social Policy
Valerie Holt, ECDataWorks
Monica King, ADRF Network
Julia Manzella, U.S. Census Bureau
Ed Mulrow, NORC at the University of Chicago
Amy O'Hara, Georgetown University
Ken Steif, University of Pennsylvania
Ali Whitmer, Georgetown University
Fred Wulczyn, Chapin Hall at the University of Chicago

CONFERENCE CO-CONVENERS

Actionable Intelligence for Social Policy (AISP), University of Pennsylvania
Actionable Intelligence for Social Policy (AISP) at the University of Pennsylvania is an initiative that focuses on the development, use, and innovation of integrated data systems (IDS) for policy analysis and program reform. AISP engages a network of states and counties that are already demonstrating the immense potential of linked administrative data to improve services and inform decision-making, as well as learning communities of developing IDS sites. By fostering collaboration, documenting best practices, and offering individualized training and technical assistance, AISP encourages social innovation and social policy experimentation so government can work better, smarter and faster.

Massive Data Institute at Georgetown University
Georgetown's main campus houses six schools that offer an array of intellectually rigorous programs designed to guide our undergraduate and graduate students toward their academic and professional goals. Our degrees include the arts and sciences, international relations, business administration, public policy, specialized certificates and continuing education programs, and executive education.

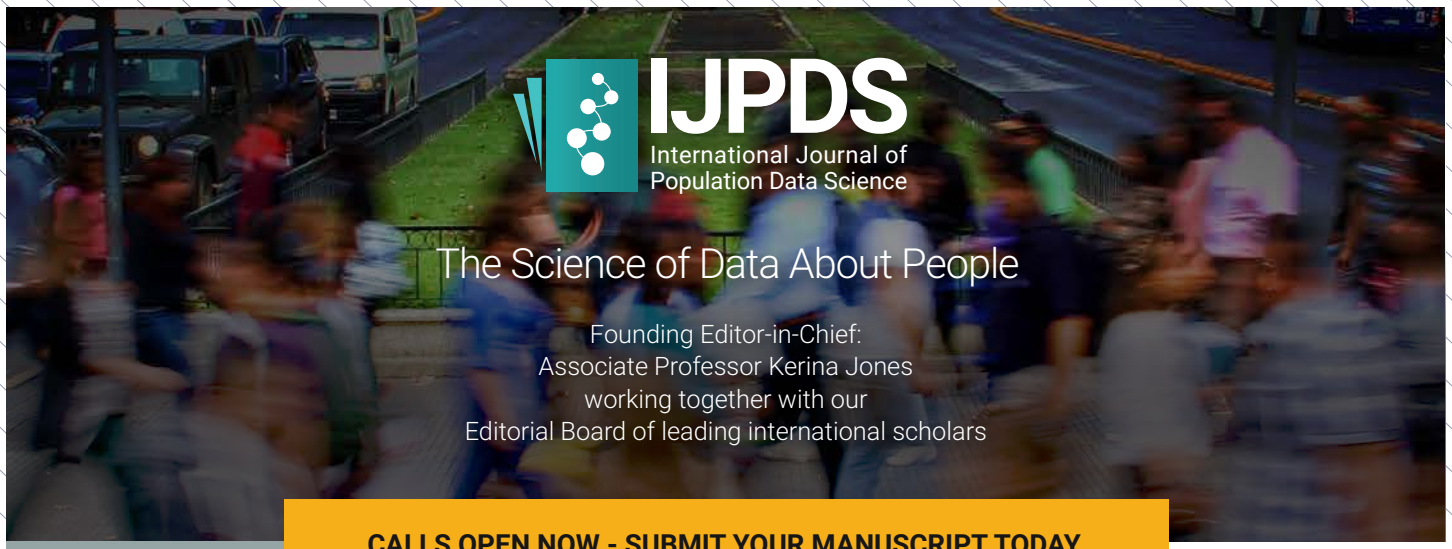
SPONSORS

Alfred P. Sloan Foundation
The Alfred P. Sloan Foundation is a philanthropic, not-for-profit grantmaking institution based in New York City. Established in 1934, the Foundation makes grants in support of original research and education in science, technology, engineering, mathematics, and economics. This conference is supported through the Foundation's Economic Institutions, Behavior & Performance program, which supports research on U.S. economic structure, behavior, and performance whose findings inform and strengthen decision-making by regulators, policymakers, and the public.

Ewing Marion Kauffman Foundation
The Ewing Marion Kauffman Foundation is a private, nonpartisan foundation that works together with communities in education and entrepreneurship to create uncommon solutions and empower people to shape their futures and be successful. The Kauffman Foundation is based in Kansas City, Missouri, and uses its $2 billion in assets to collaboratively help people be self-sufficient, productive citizens. For more information, visit www.kauffman.org and connect with us at www.twitter.com/kauffmanfdn and at www.facebook.com/kauffmanfdn.

PUBLISHER

International Journal of Population Data Science
The International Journal of Population Data Science (IJPDS) is proud to be the official publisher for the 2018 ADRF Network Research Conference. IJPDS is an open access, digital, peer-reviewed journal focusing on the science pertaining to population data. It is pioneering the way as the only journal dedicated to all aspects of Population Data Science research, development and evaluation and brings together research from multiple disciplines that would otherwise operate independently, into the field. All abstracts from the oral presentations of this conference will be published in IJPDS and made available to all via Open Access.



IJPDS

International Journal of Population Data Science

The Science of Data About People

Founding Editor-in-Chief:
Associate Professor Kerina Jones
working together with our
Editorial Board of leading international scholars

CALLS OPEN NOW - SUBMIT YOUR MANUSCRIPT TODAY

The IJPDS is an electronic, open-access, peer-reviewed journal focussing on the science pertaining to population data.

It publishes articles on all aspects of research, development and evaluation connected with data about people and populations.

- Accessing distributed data
- Analytical advances
- Architectures and infrastructures
- Capacity building
- Delivering and measuring impact
- Data and linkage quality
- Epidemiology
- Ethical, legal and societal implications (ELSI)
- Legal and regulatory issues
- Linking to emerging/complex data types
- Outcomes-based research
- Privacy-protection methodologies
- Public involvement and engagement
- Service evaluations
- Technological advances in data storage and management
- Using big data

Speed

Rapid peer review process and instant publishing as soon as manuscript is ready

Impact

Extending your impact beyond academia to the ultimate beneficiaries, the public

Reach

Established and growing international audience across the globe

Manuscripts

A broad range of manuscript formats accepted that may be out of scope for other academic journals

Find out more about IJPDS and bookmark us for your next submission: **www.IJPDS.org**

or get in touch directly: **contact@ijpds.org**

@IJPDS

linkedin.com/groups/8608483

ISSN No. 2399-4908

Publisher

Swansea University
Prifysgol Abertawe

Strategic Partner

INTERNATIONAL
Population Data Linkage
NETWORK

AGENDA *at a* GLANCE

Time	Tuesday, November 13	Wednesday, November 14
8:30–9:00am		**Breakfast and Registration** 8:30–9:00am
9:00–9:30am	**Breakfast and Registration** 9:00–9:30am	**Session C1** 9:00–10:20am / **Session D1** 9:00–10:20am
9:30–10:00am	**Remarks** 9:30–9:45am	
10:00–10:30am	**Keynote: Natalie Evans Harris** 9:45-10:30am	**Break** 10:20–10:40 am
10:30–11:00am	**Break** 10:30–10:40am	**Session C2** 10:40am–12:00pm / **Session D2** 10:40am–12:00pm
11:00–11:30am	**Session A1** 10:40am–12:00pm / **Session B1** 10:40am–12:00pm	
11:30–12:00pm		
12:00–12:30pm	**Networking Lunch** 12:00–1:00pm	**Networking Lunch** 12:00–1:30pm
12:30–1:00pm		
1:00–1:30pm	**Session A2** 1:00–2:20pm / **Session B2** 1:00–2:20pm	
1:30–2:00pm		**Best Paper Awards** 1:30–1:40pm
2:00–2:30pm		**Rapid Fire Talks** 1:40–2:30pm
2:30–3:00pm	**Coffee Break and Poster Session** 2:20pm–3:30pm	**Break** 2:30–2:45pm
3:00–3:30pm		**Closing Roundtable: Four Amys** 2:45–3:45pm
3:30–4:00pm	**Session A3** 3:30–4:50pm / **Session B3** 3:30–4:50pm	**Closing Remarks** 3:45–4:00pm
4:00–4:30pm		
4:30–5:00pm	**Break** 4:50–5:00pm	
5:00–5:30pm	**Welcome Reception** 5:00–6:00pm	
5:30–6:00pm		

ROOM LEGEND
Salon D/E
Grand III/IV
Capitol Foyer

CONFERENCE DESCRIPTION

Welcome to the 2nd annual ADRF Network Research Conference hosted by University of Pennsylvania's Actionable Intelligence for Social Policy and Georgetown University's Massive Data Institute.

Our 2018 conference builds on last year's successful inaugural event which drew nearly 150 participants from academia, government, the private sector, think tanks, and other organizations to advance administrative data use in social science research. This year, the conference program has been expanded to two full days with the theme of innovations in administrative data for social science.

Day 1 of the conference is centered on solutions to overcome technical, organizational, and methodological challenges in administrative data access and use. Sessions on Day 2 focus on examples where administrative data have been for policymaking and decision support in major domain areas.

Throughout the two days, the conference will serve as a forum to share groundbreaking work and promote interdisciplinary and cross-sector dialogue that will shape the future of the social sciences.

Tweet about the conference using #adrfconference and follow us @ADRF_Network

Tuesday, November 13, 2018

DAY 1 SCHEDULE

9:00 – 9:30
Grand III/IV

Registration and Breakfast

9:30 – 9:45
Grand III/IV

Introductory Remarks
Robert Groves, Georgetown University
Dennis Culhane, University of Pennsylvania
Monica King, University of Pennsylvania

9:45 – 10:30
Grand III/IV

Keynote Speaker: Natalie Evans Harris
Natalie has dedicated nearly 20 years to advancing the public sector's responsible use of data for protecting and improving individual's lives.

Following a 15 year career at the National Security Agency, Natalie co-founded and became COO of BrightHive, a collaborative data platform delivering a suite of smart data collection, integration, and governance products to social services providers for improved access and usability of social sector data. She founded the Community-driven Principles for Ethical Data Sharing (CPEDS) community of practice with over 800 active members focused on strengthening ethical practices in the data science community and as a Senior Policy Advisor under the Obama Administration, she founded the Data Cabinet - a federal data science community of practice with over 200 active members across more than 40 federal agencies.

Data-Driven Collective Impact: Driving Social Change as a Community
The last decade (or more) has experienced a transformation of data to an action-oriented asset that can draw insights necessary to describe, detect, predict, and evaluate factors to help our communities and the individuals in them to thrive. We've also witnessed threats to these opportunities in the forms of breaches, misinformation, and other erosions of trust that make access to and use of data much more complicated. As a community, it is imperative to take an interdisciplinary approach to data use grounded in public-private collaboration and focused on building trust with the communities we seek to serve.

10:30 – 10:40

Break

10:40 – 12:00
Salon D/E

A1: Systems Architecture, Data Processing, and Data Security
Moderator: Ken Steif (University of Pennsylvania)

PRESENTATIONS
Amy Hawn Nelson (AISP, University of Pennsylvania), Understanding the Information Architecture for Civic Innovation, Research, and Evaluation
Ryan Mackenzie White (Statistics Canada), Administrative Data Format Standardization for Efficient Analytics
Christian Hirsch (Deutsche Bundesbank), International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA)
Graham MacDonald (Urban Institute), Spark for Social Science
David Wu (Stanford University), Protecting Patient Privacy in Genomic Analysis

Tuesday, November 13, 2018

10:40 – 12:00, continued
Grand III/IV

B1: Emerging Tools, Resources, and Approaches
Moderator: Kathy Pettit (Urban Institute)

PRESENTATIONS
Kevin J Sweeney (Stats NZ), A Steady States Data Flow Model to Support Administrative Data Sourcing [Pre-recorded presentation]
Allison R.B. Tyler (University of Michigan School of Information), Researcher Credentialing for Administrative Data: Easing the Burden, Increasing the Efficiency
Robert Gradeck (University of Pittsburgh), Librarians as Critical Infrastructure in Data Ecosystem
Amanda Davis (Baltimore Neighborhood Indicators Alliance - Jacob France Institute/Enterprise Community Partners)

12:00 – 1:00
Capitol Foyer

Networking Lunch (provided)

1:00 – 2:20
Salon D/E

A2: Improving Data Quality and Standards
Moderator: Jennifer Madans (National Center for Health Statistics)

PRESENTATIONS
Zachary H. Seeskin (NORC at the University of Chicago), Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data
Paul Wormeli (Wormeli Consulting, LLC), The National Information Exchange Model and making data relevant
Sadaf Asrar (Optimal Solutions Group, LLC), A Systems Approach to International Education: How Administrative Data is Being used to Track the Progress of USAID's Education Strategy
Annette Christy (University of South Florida), Curation and Application of Statewide Data on Involuntary Assessment and Treatment in Behavioral Health
John Loft (RTI International), Electronic Health Records in Behavioral Research

Grand III/IV

B2: Measuring the Economy, Housing, and People
Moderator: Kitty Smith Evans (American Economic Association)

PRESENTATIONS
Marvin Ward Jr. (JPMorgan Chase Institute), Our Economy is Evolving: Shouldn't the Way We Measure It Evolve Too?
Carla Medalia (U.S. Census Bureau), Linking Survey and Administrative Data to Measure Income, Inequality, and Mobility
John Haltiwanger (University of Maryland), Minding Your Ps and Qs: Going from Micro to Macro in Measuring Prices and Quantities
Misty L. Heggeness (U.S. Census Bureau), Harnessing Administrative Records for Official Statistics on People and Households
Laurie Goodman (Urban Institute), Housing Affordability: Local and National Perspectives

Tuesday, November 13, 2018

DAY 1 SCHEDULE

2:20 – 3:30
Capitol Foyer

Coffee Break, Poster, and Interactive Session

POSTERS AND INTERACTIVE SESSION
Kara Bonneau (Duke University), North Carolina Education Research Data Center
Megan Davis (Mathematica Policy Research), Improving Administrative Data Quality at the Initial Stages of the Data Life Cycle
Sandra Clark (U.S. Census Bureau), Using Administrative Records in the American Community Survey
David Bleckley (University of Michigan/ICPSR), DataLumos and Beyond: ICPSR's Tool for Crowdsourced Sharing of Administrative and Other Government Data and Ways to Address its Limitations
Brian J. Goode (SDAL - Virginia Tech), Classifying Families and Households for VA Government Stakeholders using Existing Administrative Records
Misty Heggeness (US Census Bureau), Linking Administrative Data across Federal Agencies: Outcome Evaluations for NIH Training Programs
Jordan Butz (University of Pennsylvania) and Annie Streetman (University of Pennsylvania), Predicting Spatial Risk of Opioid Overdoses in Providence, RI
Kathryn Shantz (U.S. Census Bureau), Using SNAP and TANF Administrative Records and the Transfer Income Model to Evaluate Poverty Measurement
Adela Luque (U.S. Census Bureau), The Nonemployer Statistics by Demographics: Leveraging Administrative Data to Meet Stakeholder Needs
Susan Jekielek (University of Michigan), Higher Education Administrative Data Elements: Potential for Analysis
Amy O'Hara (Georgetown University), "I Need That!" Interactive Session

3:30 – 4:50
Salon D/E

A3: Innovations in Data Linkage
Moderator: Luiza Antonie (University of Guelph)

PRESENTATIONS
Ian Crandell (Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech), Record Linkage Reconciliation of Arlington Department of Human Services Administrative Data Using Potts Models
Dean M. Resnick (NORC), Simulation Approach to Assess the Precision of Estimates Derived from Linking Survey and Administrative Records
Marc Roemer (NCHS), An assessment of using frequency weights for record linkage
Susan Hautaniemi Leonard (ICPSR/University of Michigan), Building a repository for record linkage

Grand III/IV

B3: Recipes for Successful Collaborations Across Organizations
Moderator: Della Jenkins (Actionable Intelligence for Social Policy)

PRESENTATIONS
Heather L. Rouse (Iowa State University), Integrated Administrative Data for Early Childhood Iowa: A Governance Model to inform Policy and Program Collaboration
James Ayles (New Brunswick Department of Health), Key Factors in the establishment an academia-government center of public sector administrative data and policy research
Jennifer Romich (University of Washington), The Washington State Merged Longitudinal Administrative Database
Jennifer Auer (Optimal Solutions Group, LLC), Working with federal government agencies to unlock administrative data
Roxane Silberman (CASD-GENES & CNRS Paris, France), International Data Access Network (IDAN)

4:50 – 5:00

Break

5:00 – 6:00
Capitol Foyer

Welcome Reception
Sponsored by the Stone Chair in Social Policy at the University of Pennsylvania

Wednesday, November 14, 2018

DAY 2 SCHEDULE

8:30 – 9:00
Grand III/IV

Registration and Breakfast

9:00 – 10:20
Salon D/E

C1: Uses of Administrative Data for Supporting Education Policy
Moderator: Katie Barghaus (University of Pennsylvania)

PRESENTATIONS
Nikolas Pharris-Ciurej (US Census Bureau), Demonstrating the Use of Linked Data to Advance Education Research
Douglas Lauen (University of North Carolina at Chapel Hill), Early College High Schools at Scale: Using Administrative Data to Assess the Impacts of an Educational Intervention on Voting and Crime
Sally Wallace (Georgia State University), Returns to Late Aged College Degrees
Wesley Greenblatt (MIT), Long-Term Effects from Early Exposure to Research: Evidence from the NIH "Yellow Berets"

Grand III/IV

D1: Uses of Administrative Data for Supporting Public Programs and Public Health Policy
Moderator: Kelly Bidwell (Office of Evaluation Sciences, GSA)

PRESENTATIONS
Stephanie Shipp (Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia), Developing a Data-driven Approach to inform Planning in County Health and Human Services Departments in the Context of a Case Study on Obesity
Jenny Povey (The University of Queensland), The power of linked data: Evaluating diverse multi-program projects designed to reduce welfare dependence
Alex Collie (Monash University Australia), Evaluating the impact of workers' compensation policy in Australia using insurance claims data and comparative quasi-experimental methods
Rachel Shattuck (U.S. Census Bureau), The Use of Blended Data to Improve Public Assistance Programs: Results from a Partnership between the U.S. Census Bureau, USDA, and State Program Agencies
Kevin P. Conway (RTI International), Fusing Administrative Data to Combat the Opioid Crisis

10:20 – 10:40

Break

10:40 – 12:00
Salon D/E

C2: Uses of Administrative Data for Supporting Child Welfare Policy
Moderator: Kristen Coe (University of Pennsylvania)

PRESENTATIONS
Francisca Richter (Case Western Reserve University), Using Integrated Data to Design and Support Pay For Success Interventions
Ken Steif (University of Pennsylvania), Developing a spatial risk prediction model for child maltreatment
Dyann Daley (Predict-Align-Prevent, Inc.), Policy and resource optimization based on spatial risk prediction models for child maltreatment
Jessica Raithel (Center for Innovation through Data Intelligence), Typologies of Transition-Age Youth
Angela You Gwaltney (University of North Carolina, Chapel Hill), Effect of Psychotropic Medication on Foster Care Experience and Outcomes: A Causal Analysis using Administrative Data

DAY 2 | WEDNESDAY, NOVEMBER 14, 2018

10:40 – 12:00, continued
Grand III/IV

D2: Other Uses of Administrative Data to improve Decision-making and Program Evaluation
Moderator: Nick Hart (Bipartisan Policy Center)

PRESENTATIONS
Joshua Goldstein (Social and Decision Analytics Division, Biocomplexity Institute of University of Virginia)
Natsuko Nicholls (Institute for Research on Innovation and Science (IRIS) / Univ of Michigan Institute for Social Research), Use and Application of Federal Advisory Committee Act (FACA) Database
Sadaf Asrar (Optimal Solutions Group, LLC), Stretching a Buck: Using Administrative Data to Inform Continuous Quality Improvement
Marden F. Munoz (The Children's Trust), Trust Central Eases Funding Decisions With Data
Francis Mitrou (The University of Western Australia), Aboriginal life pathways through multiple human service domains; administrative data linkage for policy

12:00 – 1:30
Capitol Foyer

Networking Lunch (provided)

1:30 – 1:40
Grand III/IV

Presentation of the Best Paper Awards

1:40 – 2:30
Grand III/IV

Rapid Fire Talks
Moderator: Tanvi Desai (Data Policy and Strategy Consultant)

PRESENTATIONS
Michael Lenczner (Powered by Data), Overcoming policy barriers to administrative data sharing through an inclusive civil society coalition
Christian Hirsch (Deutsche Bundesbank), The Role of the Bundesbank Microdata Production in Times of Big Data: The Need for Data Access, Data Sharing and for an Integrated Digital Information System
Lars Vilhuber (Cornell University), Reproducible research in administrative data center
Sarah Strochak (Urban Institute), Is Limited English Proficiency a Barrier to Homeownership?
Benoit Dostie (HEC Montreal), Immigrant Careers and Networks
Brittany Borg (US Small Business Administration), Program Evaluation Using Multiple Datasets
Nikolas Pharris-Ciurej (Census Bureau), Blended data: a novel opportunity to advance survey operations and knowledge of the US economy and population
Aaron D. Schroeder (Social and Decision Analytics Division, Biocomplexity Institute of University of Virginia), Complying with New Address Requirements for the 2020 Census: Using Open Administrative Records and Delivery Point Validation Data to Discover Verified Multi-Family Unit Addresses in Arlington County, Virginia
Rachel Carnahan (NORC at the University of Chicago), Using Administrative Data to Reduce Respondent Burden in Facility Data Collection

2:30 – 2:45

Break

2:45 – 3:45
Grand III/IV

Closing Roundtable: Four Amys
In this panel, four Amys will discuss how they have handled challenges involving administrative data access. These Amys have needed data and pleaded for data. They have been data owners, linkers, gatekeepers, and analysts. Michael Hawes, Director of Student Privacy Policy at the U.S. Department of Education will moderate the panel focusing on approaches to overcome operational and organizational challenges.

Amy Hawn Nelson is the director for training and technical assistance at AISP, overseeing their Learning Communities Initiative. Prior to joining AISP in 2017, Amy Hawn Nelson was the director of Social Research for the UNC Charlotte Urban Institute, and director of the Institute for Social Capital, the Integrated Data Systems in the Charlotte region.

Amy Laitinen is the director for Education Policy at New America, focusing on policies to increase quality and transparency in higher education. Amy Laitinen previously served as a policy advisor on higher education at both the U.S. Department of Education and the White House.

Amy Nowell is chair of the Institutional Review Board (IRB) at MDRC, considering the ethical use of old RCT studies among other data privacy issues. Amy Nowell was previously director of Research and Evaluation and External Research Coordination at Chicago Public Schools. outcomes.

Amy O'Hara works in Georgetown University's Massive Data Institute, developing administrative and technical capacity to improve data access for social science research and program evaluation. Before coming to Georgetown, Amy led the U.S. Census Bureau's Center for Administrative Records Research and Applications.

Moderator: Michael Hawes (Department of Education)

PANELISTS
Amy Hawn Nelson (AISP, University of Pennsylvania)
Amy Laitinen (New America)
Amy Nowell (MDRC)
Amy O'Hara (Stanford University)

3:45 – 4:00
Grand III/IV

Closing Remarks

DAY 1 ABSTRACTS

10:40 – 12:00
A1: SYSTEMS ARCHITECTURE, DATA PROCESSING, AND DATA SECURITY

Understanding the Information Architecture for Civic Innovation, Research, and Evaluation

- Natalie Evans Harris (BrightHive); Amy Hawn Nelson (Actionable Intelligence for Social Policy, University of Pennsylvania)

  Efforts to collect, manage, transform, and integrate data across administrative systems into actionable knowledge to inform better policy decisions are becoming more common. However, the technical processes, procedures, and infrastructure they employ vary substantially. Variety in approaching data infrastructure, transfer, linking, and security is expected in this emerging field, but both established and developing efforts would benefit from cohesive guidance regarding the technical considerations of data integration, with focus on presenting a range of options that can be weighted based on context specific restrictions (e.g. cost, staffing, or existing infrastructure).

  Actionable Intelligence for Social Policy (AISP), MetroLab Network, and the National Neighborhood Indicators Partnership (NNIP) with support from the Annie E. Casey Foundation, are convening a working group to shape and develop guidance on information architecture and technical approaches for data integration efforts such as those in the AISP and NNIP networks and the AISP Learning Community. This guidance will help newly emerging efforts as well as established ones looking to update their current approach. It will also inform policymakers and researchers who need a primer to better understand the technical components and considerations at play for data sharing and integration. This presentation will present findings, best practices and recommendations from this brief that will be released in Fall 2018.

International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA)

- Christian Hirsch (Deutsche Bundesbank)

  The financial crisis of 2007/08 has highlighted the need for using granular data on financial institutions and markets to detect risks and imbalances in the financial sector. Administrative data producers are witnessing a growing need to improve granular data access and sharing. When sharing granular data, data producers face significant legal and technical challenges related to, among others, safeguarding statistical confidentiality.

  To meet the demand of data users and data compilers for (granular) data sharing, the International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA) was established on 6 January 2017. Current INEXDA members are the Banco de España, Banca d'Italia, the Banco de Portugal, the Bank of England, the Banque de France, the Deutsche Bundesbank, and the European Central Bank.

  INEXDA provides a platform for administrative data producers to exchange practical experiences on the accessibility of granular data, on metadata as well as on techniques for statistical analysis and data protection. During the INEXDA meeting in Paris, Julia Lane has presented the Administrative Data Research Facility (ADRF). Following her presentation several INEXDA members expressed interest in obtaining more information about the approach taken by ADRF. This presentation describes the interaction between INEXDA and ADRF.

Spark for Social Science

- Graham MacDonald (Urban Institute); Alex Engler (University of Chicago); Jeffrey Levy (Urban Institute); Sarah Armstrong (University of Chicago)

  Urban has developed an elastic and powerful approach to the analysis of massive datasets using Amazon Web Services' Elastic MapReduce (EMR) and the Spark framework for distributed memory and processing. The goal of the project is to deliver powerful and elastic Spark clusters to researchers and data analysts with as little setup time and effort possible, and at low cost. To do that, at the Urban Institute, we use two critical components: (1) an Amazon Web Services (AWS) CloudFormation script to launch AWS Elastic MapReduce (EMR) clusters (2) a bootstrap script that runs on the Master node of the new cluster to install statistical programs and development environments (RStudio and Jupyter Notebooks). The Urban Institute's Spark for Social Science Github page holds code used to setup the cluster and tutorials for learning how to program in R and Python.

Administrative Data Format Standardization for Efficient Analytics

- Ryan Mackenzie White **(Statistics Canada)**

Adoption of non-traditional data sources to augment or replace traditional survey vehicles can reduce respondent burden, provide more timely information for policy makers, and gain insights into the society that may otherwise be hidden or missed through traditional survey vehicles. The use of non-traditional data sources imposes several technological challenges due to the volume, velocity and quality of the data. The lack of applied industry-standard data format is a limiting factor which affects the reception, processing and analysis of these data sources. The adoption of a standardized, cross-language, in-memory data format that is organized for efficient analytic operations on modern hardware as a system of record for all administrative data sources has several implications:

- Enables the efficient use of computational resources related to I/O, processing and storage.
- Improves data sharing, management and governance capabilities.
- Increases analyst accessibility to tools, technologies and methods.

Statistics Canada developed a framework for selecting computing architecture models for efficient data processing based on benchmark data pipelines representative of common administrative data processes. The data pipelines demonstrate the benefits of a standardized data format for data management, and the efficient use of computational resources. The data pipelines define the preprocessing requirements, data ingestion, data conversion, and metadata modeling, for integration into a common computing architecture. The integration of a standardized data format into a distributed data processing framework based on container technologies is discussed as a general technique to process large volumes of administrative data.

Protecting Patient Privacy in Genomic Analysis

- David Wu **(Stanford University)**

Patient genomes are interpretable only in the context of other genomes. However, privacy concerns over genetic data oftentimes deter individuals from contributing their genomes to scientific studies and prevent researchers from sharing their data with the scientific community. In this talk, I will describe how we can leverage secure multiparty computation techniques from modern cryptography to perform useful scientific computations on genomic data while protecting the privacy of the participants' genomes. In multiple real scenarios, our methods successfully identified the disease-causing genes and even discovered previously unrecognized disease genes, all while keeping nearly all of the participants' most sensitive genomic information private. We believe that our techniques will help make currently restricted data more readily available to the scientific community and enable individuals to contribute their genomes to a study without compromising their personal privacy.

The material from this talk is based on joint works with Gill Bejerano, Bonnie Berger, Johannes A. Birgmeier, Dan Boneh, Hyunghoon Cho, and Karthik A. Jagadeesh.

10:40 – 12:00
B1: EMERGING TOOLS, RESOURCES, AND APPROACHES

Utilizing data-driven technology tools for community-led solutions to vacant properties and urban blight

- Amanda M. Davis **(Baltimore Neighborhood Indicators Alliance - Jacob France Institute/Enterprise Community Partners);** Seema D. Iyer **(Baltimore Neighborhood Indicators Alliance, The Jacob France Institute, University of Baltimore);** Kristine J. Dunkerton **(Community Law Center);** Shana Roth-Gormley **(Community Law Center)**

As a city that has lost more than 1/3 of its population over the past 6 decades, some Baltimore neighborhoods suffer from a disproportionate number of vacant and abandoned properties, mired in issues of unclear ownership and "under-water" lien burdens. Cloudy legal and financial restrictions cause properties to cycle through a speculative system that strips them of all equity, and causes them to move out of reach for redevelopment. Evidence suggests that existing processes for addressing these issues, such as tax lien sales and foreclosures, can actually play a role in increasing vacancy rates and amplify neighborhood disinvestment (Dewar, Seymour, and Druță, 2015). Policies aimed at real property tax reform and foreclosure prevention can improve conditions, yet communities, non-profits, and city agencies in Baltimore lacked a unified data system to guide their reform and outreach efforts.

One challenge is that property data are housed at various agencies, each using its own system of data storage and dissemination, making it difficult to use different datasets for a single property. The Baltimore City Open Land Data (BOLD) web application arose out of the need to streamline the data gathering process by integrating various datasets for easier use by stakeholders working to stabilize their communities, preserve homeownership, and break the cycle of vacant properties. This presentation will give an overview how BOLD was designed, a short demonstration of the application, and show how it can be used to further research the impact of tax sales and foreclosures in Baltimore City.

Librarians as Critical Infrastructure in Data Ecosystems

- Robert Gradeck **(University of Pittsburgh)**

Across the US and around the world, a growing number of public sector, educational, and nonprofit organizations have been sharing data with one another. These organizations hope to increase transparency, enhance efficiency and service quality, improve communities, encourage public participation, develop new knowledge, and foster civic innovation. While there are many success stories around data sharing, there is a growing awareness that the act of publishing data will not always result in community impact. Data intermediaries are often needed to help people extract value from data, and to help producers make good decisions about what and how they publish.

In Pittsburgh, our local civic data ecosystem is unique in that both public and academic librarians are actively involved as data intermediaries, and they work in close collaboration with other intermediaries, data publishers and users in a variety of ways. Librarians play a number of roles, including helping people discover information, building data literacy and technical skills, providing technical assistance in data management and documentation, creating feedback mechanisms to publishers, convening and hosting events, and connecting data users.

Our experience shows that libraries and librarians should be key actors in the continuing development of data ecosystems and act as core data intermediaries; their expertise adds value to a wide range of issues that affect both data publishers and users. In this talk, I will share insights gained through the Civic Switchboard project, which aims to develop the capacity of libraries in civic data ecosystems.

A Steady States Data Flow Model to Support Administrative Data Sourcing

- Kevin J Sweeney **(Stats NZ)**

Administrative data represents the future sourcing status quo for government social sector agencies, increasingly replacing more expensive, labour-intensive and less timely survey collection.

But an administrative data model brings challenges, based in the inescapable fact that such data has been collected by a different organisation, for a different purpose. The resultant cost to unpack undesirable qualities in the data can be significant and absorb inordinate levels of energy better spent deriving actionable insights. If the administrative data is integrated with other data, these costs are multiplied.

New Zealand' national statistical agency has developed a new style of data governance framework that includes a steady state model (SSM) for mapping data flow. As innovation, the SSM offers an effective way to articulate administrative data issues, supporting improved supplier-user coordination. As such it can potentially reduce data handling costs, bolstering the value proposition for administrative data generally.

By providing a means of reflecting user organisation data lifecycle perspectives within the supplier' business process environment, the SSM opens an engagement channel that bridges what can otherwise represent insurmountable communication barriers. And since both user and supplier are able to maintain their native frame of reference, the resulting engagement fosters collaboration while facilitating what is more likely to represent a mutually amenable outcome.

Researcher Credentialing for Administrative Data: Easing the Burden, Increasing the Efficiency

- Allison R.B. Tyler **(University of Michigan School of Information)**; Johanna Davidson Bleckman **(Inter-university Consortium for Political and Social Research)**; Margaret C. Levenstein **(Inter-university Consortium for Political and Social Research)**

While not collected primarily for research purposes, administrative data gathered as part of normal agency or program operations present unique and vital opportunities for researchers. However, as with other types of restricted

access data, these data are often not made available or require special authorization to access. The risks of re-identification, social stigma, and privacy violations for individuals represented in the data, especially for special populations, require that data be securely held and access authorization be moderated. Obtaining this authorization, especially for multiple datasets from one institution or spread across multiple data providers, imposes a significant administrative burden on researchers and data repository staff who must repeatedly and redundantly provide and validate user identities.

This presentation will offer the Inter-university Consortium for Political and Social Research's solution to ease the burden on both researchers and data repository staff -- the Researcher Passport. The researcher passport will incorporate standardized, community-normed identity verification criteria, data security level interpretations, and restricted data access training requirements. The researcher passport, using these standards, will be issued by a central authority who carries out the identity verification process and issues a tiered-access passport to users who are authorized for a streamlined data access request process for certain levels of secure data. Visas issued by data custodians control "entry" by passport holders to particular data sets. We describe these standards, how they fit into the repository workflow, and how they make the data access process for efficient and effective for administrative data users and providers.

1:00 – 2:20
A2: IMPROVING DATA QUALITY AND STANDARDS

A Systems Approach to International Education: How Administrative Data is Being used to Track the Progress of USAID's Education Strategy

- Sadaf Asrar **(Optimal Solutions Group, LLC)**

  While rigorous evaluation of donor funded education interventions are quite common in developing countries, the data collected and analyzed to assess the interventions are not always published. Even when published, the data often lack codebooks and technical documentation, guidance on how to correctly use the data, and most frustratingly only contain meta-information pertaining to the design of the intervention and evaluation in free text in PDF reports. Such practices reduce the overall quality of the data and meta-data making it very difficult to be used for secondary analysis to answer policy relevant questions. To address such challenges, this paper describes the integrated data system developed to securely receive, review, harmonize, ingest, curate, and publish standardized public use Early Grade Reading Assessment data and meta-data to track the progress of the 2011-2015 USAID Education Strategy. The paper discusses the underlying methodological framework for systematically reviewing the data for quality and completeness as well as the approach to profile the design of the intervention and evaluation to assess its rigor using the system. Lastly, the paper illustrates how the meta-data created through the systematic review process combined with the data collected and processed through the system can be used for developing dynamic data products that reduce the technical barriers for such data to be used by a wide range of stakeholders to inform critical policy decisions.

Data Curation and Application of Statewide Data on Involuntary Assessment and Treatment in Behavioral Health

- Annette Christy **(University of South Florida)**; Daniel Ringhoff **(University of South Florida)**; Sara Rhode **(University of South Florida)**; Paige Alitz **(University of South Florida)**

  Our university, based data center has curated statewide data on short-term, involuntary examinations for mental illness/co-occurring disorders for over two decades. We recently began receiving petitions and orders for longer-term civil commitment from Clerks of Court. We are currently developing a system to curate data on involuntary assessments/treatment for substance use disorders. The involuntary examination data have been used to produce 100+ ad hoc reports for a variety of stakeholders, a statutorily required annual report, as well as to inform the state legislature, advocates, agencies, and several statewide taskforces relating to criminal justice and mental health initiatives. Data from documents are currently received and entered in a) hard copy via the mail, b) securely scanned and transferred either via SFTP or with secure transfer to our University's Box.com account, or c) direct provider entry into a secure web portal. Our University's IT environment has evolved, with an escalation of organizational and policy changes related to the merging of two IT units. While this merging has led to innovation, it has also presented operational, organizational and logistical challenges. Discussed in this presentation will be a) these IT challenges, b) the pros and cons of form submission methods, c) how choice of submission method is informed by the capabilities of those submitting the documents in addition to, the resources and capabilities of our center within the context of current funding, as well as d) how this impacts choices made about data entry, data quality and use of the data for analyses.

Electronic Health Records in Behavioral Research

- John Loft (RTI International); Diana Greene (RTI International)

Medical records are a type of administrative record with rich potential for research of behavioral health and health policy. Developments in electronic health records (EHR) can increase access to data contained in medical records but also present some unusual challenges for research. This presentation summarizes recent literature describing the use of EHR in research and identifies issues for consideration in the preparation of research design and protocols for data collection and preparation. The discussion is presented in a framework for evaluation of data quality and fitness for use.

Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data

- Zachary H. Seeskin (NORC at the University of Chicago); A. Rupa Datta (NORC at the University of Chicago); Gabriel Ugarte (NORC at the University of Chicago)

State and local agencies administering programs have in their administrative data a powerful resource for policy analysis to inform evaluation and guide improvement of their programs. Understanding different aspects of their administrative data quality is critical for agencies to conduct such analyses and to improve their data for future use. However, state and local agencies often lack the resources and training for staff to conduct rigorous evaluations of data quality. We describe our efforts developing tools that can be used to assess data quality as well as the challenges encountered in constructing these tools. The toolkit focuses on critical dimensions of quality for analyzing an administrative dataset, including checks on data accuracy, the completeness of the records, and the comparability of the data over time and among subgroups of interest. State and local administrative databases often include a longitudinal component which our toolkit also aims to exploit to help evaluate data quality. While we seek to develop general tools for common data quality analyses, most administrative datasets have particularities that can benefit from a customized analysis building on our toolkit. In addition, we incorporate data visualization to draw attention to sets of records or variables that contain outliers or for which quality may be a concern.

The National Information Exchange Model and making data relevant

- Paul Wormeli (Wormeli Consulting, LLC; Adjunct professor, George Washington University)

The National Information Exchange Model (NIEM) has been widely adopted as a basis for creating standards that foster information exchanges across disparate disciplines in justice, health, human services and other fields. NIEM can be a driver for transforming legacy data into meaningful cross-disciplinary exchanges that can support advanced analysis of impacts of programs and strategies in improving government services of all categories. This presentation will review the basic NIEM principles in terms of how they can enable more meaningful results as administrative data is applied to research on the efficacy of government services across organizations and disciplines.

1:00 – 2:20
B2: MEASURING THE ECONOMY, HOUSING, AND PEOPLE

- Gabe Ehrlich (University of Michigan); John Haltiwanger (University of Maryland); Ron Jarmin (Bureau of the Census); David Johnson (University of Michigan)

Abstract Content: Key macro indicators such as output, productivity and inflation are based on a complex system of collection from different samples and different levels of aggregation across multiple statistical agencies. The Census Bureau collects nominal sales, the Bureau of Labor Statistics collects prices, and the Bureau of Economic Analysis constructs nominal and real GDP using these and other data sources. The price and quantity data are integrated at a high level of aggregation (product and industry classes). A similar mismatch of price and nominal variables pervades the productivity data, which use industry-level producer price indexes as deflators. This paper explores alternative methods for re-engineering key national output and price indices using transactions-level data. Such re-engineering offers the promise of greatly improved macroeconomic data along many dimensions. First, price and quantity would be based on the same observations. Second, the granularity of data could be greatly increased on many dimensions. Third, time series could be constructed at a higher frequency and

on a more timely basis. Fourth, the use of transactions-level data opens the door to new methods for tracking product turnover and other sources of product quality change that may be biasing the key national indicators. Implementing such a new architecture for measuring economic activity and price change poses considerable challenges. This paper explores these challenges, along with a re-engineered approach's implications for the biases in the traditional approaches to measuring output growth, productivity growth, and inflation.

Harnessing Administrative Records for Official Statistics on People and Households

- Misty L. Heggeness **(U.S. Census Bureau)**

The availability and excessiveness of alternative (non-survey) data sources, collected on a daily, hourly, and sometimes second-by-second basis, has challenged the federal statistical system to update existing protocol for developing official statistics. Federal statistical agencies collect data primarily through survey methodologies built on frames constructed from administrative records. They compute survey weights to adjust for non-response and unequal sampling probabilities, impute answers for nonresponse, and report official statistics via tabulations from these survey. The U.S. federal government has rigorously developed these methodologies since the advent of surveys -- an innovation produced by the urgent desire of Congress and the President to estimate annual unemployment rates of working age men during the Great Depression.

In the 1930s, Twitter did not exist; high-scale computing facilities were not abundant let alone cheap, and the ease of the ether was just a storyline from the imagination of fiction writers. Today we do have the technology, and an abundance of data, record markers, and alternative sources, which, if curated and examined properly, can help enhance official statistics. Researchers at the Census Bureau have been experimenting with administrative records in an effort to understand how these alternative data sources can improve our understanding of official statistics. Innovative projects like these have advanced our knowledge of the limitations of survey data in estimating official statistics. This paper will discuss advances made in linking administrative records to survey data to-date and will summarize the research on the impact of administrative records on official statistics.

Our Economy is Evolving: Shouldn't the Way We Measure It Evolve Too?

- Marvin Ward Jr. **(JPMorgan Chase Institute);** Bryan Kim **(JPMorgan Chase Institute);** Lindsay Relihan **(JPMorgan Chase Institute);** James Duguid **(JPMorgan Chase Institute)**

The Local Consumer Commerce Index is a measure of local economic activity parsed by a variety of consumer and merchant characteristics. By leveraging an administrative database of over 24 billion debit and credit card transactions made by over 64 million de-identified customers, this index from the JPMorgan Chase Institute addresses the lack of data series with sufficient spatiotemporal and demo/firmographic resolution to support tactical decision making in local economies.

Each transaction carries the age and income of the consumer, the merchant size and type of product it sells, as well as the zip code of both. Using these characteristics we construct a measure of year-over-year spending growth by consumers at merchants located in 14 major metropolitan areas in the US. The index data are screened and weighted to represent population-wide spending levels. This unique lens on local economies is freely provided to the public in accordance with the Institute's mission of advancing the public good.

We have also extended this data asset beyond its use for reporting and economic monitoring. One extension has been our research that measures intra-city demand. By measuring the distance between where consumers live and the merchants at which they shop, we have lent nuance and granularity to policy discussions surrounding intra-city inequities in economic vitality.

We hope to socialize the power of leveraging administrative data for the public good, in hopes that other administrative data-owners are encouraged to also furnish analyses based on their administrative data to help inform the public policy process.

Linking Survey and Administrative Data to Measure Income, Inequality, and Mobility

- Carla Medalia (U.S. Census Bureau); Bruce Meyer (University of Chicago); Amy O'Hara (Stanford University); Derek Wu (University of Chicago)

  Income is one of the most important measures of well-being, but it is notoriously difficult to measure accurately. Income data are available from surveys, tax records, and government programs, but each of these sources has important strengths and major limitations when used alone. We are linking multiple data sources to develop the Comprehensive Income Dataset (CID), a restricted micro-level dataset that combines the demographic detail of survey data with the accuracy of administrative measures. By incorporating information on nearly all taxable income, tax credits, and cash and in-kind government transfers, the CID surpasses previous efforts to provide an accurate and comprehensive measure of income for the population of U.S. individuals, families, and households. We use models to evaluate differences across the data sources and explore imputation methods and trends over time. The CID can enhance Census Bureau surveys and statistics through investigating measurement error, improving imputation methods, and augmenting surveys with the best possible estimates of income. It can also be used to improve the administration of taxes by the Internal Revenue Service and forecast and simulate changes in programs and taxes. Finally, the CID has substantial advantages over other sources to analyze numerous research topics, including poverty, inequality, mobility, and the distributional consequences of government transfers and taxes.

Housing Affordability: Local and National Perspectives

- Laurie Goodman (Urban Institute), Wei Li (Federal Deposit Insurance Corporation); Jun Zhu (Urban Institute)

  This paper presents a new approach to measuring affordable homeownership. Future changes in the homeownership rate will depend on the ability of today's renters to become homeowners. Our proposed housing affordability for renters index (HARI) focuses on how affordable homeownership is for current renters. We look at the share of renters who reported the same or more income than those who recently purchased a home using a mortgage, in effect measuring how many renters have enough income to purchase a house. For each metropolitan statistical area (MSA), we construct a local area index that compares renters and borrowers in the same MSA and a national index that compares renters nationwide with homeowners in a specific MSA. We rely on the Administrative Data Research Facility to construct these indices. This database, constructed by the Urban Institute, aggregates American Community Survey variables and Home Mortgage Disclosure Act variables to common geographies. The new indices reveal that slightly more than a quarter of current US renters have incomes higher than those who recently became homeowners using a mortgage. The indices also reveal how housing affordability differs over time and across race/ethnicity groups and locations. We demonstrate the value of our new indices by showing that they are predictive of homeownership rates: MSAs that are deemed more affordable by our index have higher homeownership rates.

3:30 – 4:50
A3: INNOVATIONS IN DATA LINKAGE

Record Linkage Reconciliation of Arlington Department of Human Services Administrative Data Using Potts Models

- Ian Crandell (Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech); Aaron Schroeder (Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech); Dave Higdon (Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech); Michael-dharma Irwin (Arlington County Department of Human Services Ian Crandell)

  Situated at the nexus of federal, state, and local governments, the Arlington Department of Human Services (DHS) receives service utilization data from a multitude of different sources. Because of their "no wrong door" policy, customers can sign up for any DHS service from any DHS department. A practical consequence of this is that a single person can appear as multiple records from multiple databases with no unambiguous key between these records. Merging these records requires a probabilistic linkage approach. Classical approaches to record linkage, such as the method of Felligi and Sunter, consider each possible pair of records between databases and assigning link probabilities to each one. A drawback of considering pairwise links alone is that sometimes the transitive nature of links is violated. In order to better handle such information clashes, we propose a Bayesian linkage method that considers a large set of possible pairs

at once. At the heart of this approach is a Potts model representation that tracks which records are assigned to the same individual. This allows us to assign probabilities to the various reconciliations of inconsistent linkage assignments.

Building a repository for record linkage

- Susan Hautaniemi Leonard (ICPSR/University of Michigan); Abay Israel (ICPSR/University of Michigan); Margaret Levenstein (ICPSR/University of Michigan); Trent Alexander (ICPSR/University of Michigan)

ICPSR is building LinkageLibrary, a repository and community space for researchers involved in linking and combining datasets, as a collaboration between social, statistical, and computer scientists. Unlike surveys or experiments where causal and outcome variables are measured in tandem, it is often necessary when working with organic, non-design data to link to other measures. This makes linkage methodologies particularly important when conducting analyses using administrative data. A common benchmarking repository of linkage methodologies will propel the field to the next level of rigor by facilitating comparison of different algorithms, understanding which types of algorithms work best under different conditions and problem domains, promoting transparency and replicability of research, and encouraging proper citation of methodological contributions and their resulting datasets. It will bring together the diverse scholarly communities (e.g., computer scientists, statisticians, and social, behavioral, economic, and health (SBEH) scientists) who are currently addressing these challenges in disparate ways that do not build on one another's work. Improving linkage methodologies is critical to the production of representative samples, and thus to unbiased estimates of a wide variety of social and economic phenomena. The repository will accelerate the development of new record linkage algorithms and evaluation methods, improve the reproducibility of analyses conducted on integrated data, allow comparisons on same and different data, and move forward the provision of privacy-aware integrated data. The presentation will focus on lessons learned while building the repository and the community, and introduce the LinkageLibrary website.

Simulation Approach to Assess the Precision of Estimates Derived from Linking Survey and Administrative Records

- Dean M. Resnick (NORC at the University of Chicago); Lisa B. Mirel (NCHS)

Probabilistic record linkage implies that there is some level of uncertainty related to the classification of pairs as links or non-links vis-à-vis their true match status. As record linkage is usually performed as a preliminary step to developing statistical estimates, the question then is how does this linkage uncertainty propagate to them? In this paper, we develop an approach to estimate the impact of linkage uncertainty on derived estimates by using a re-sampling approach. For each iteration of the re-sampling, pairs are classified as links or non-links by Monte-Carlo assignment to model estimated true match probabilities. By looking at the range of estimates produced in a series of re-samples, we can estimate the distribution of derived statistics under the prevailing incidence of linkage uncertainty. For this analysis we use the results of linking the 2014 National Hospital Care Survey to the National Death Index performed at the National Center for Health Statistics. We assess the precision of hospital-level death rate estimates.

An assessment of using frequency weights for record linkage

- Marc Roemer (NHCS); Scott Campbell (NORC at the University of Chicago)

Many different techniques can be used to perform entity-to-entity record linkage. The optimal approach may relate to the entity type, such as individual or establishment, and source, such as sample survey, enumeration, program administration, or register. Within the Fellegi-Sunter record linkage framework, the frequency of occurrence of match variables' values can be either employed or ignored when estimating the probabilities from which the match weights are derived. Namely, the U-probability that a variable agrees in a pair of non-matched records, is estimated for each value of a match variable in the frequency-based approach, or in general in the non-frequency-based approach. The aim of this talk is to compare the quality of results produced by the frequency-based and non-frequency-based approaches when linking a household survey and an establishment survey to vital records. A household survey, the National Health Interview Survey (NHIS) and an establishment survey, the National Hospital Care Survey (NHCS) are each linked to the National

Death Index (NDI) using the Fellegi-Sunter record linkage framework. We perform the linkage on each survey twice; first, employing frequency-based weights for all match variables, and second, simple agree/disagree weights for all match variables. We then examine any differences in quality within each survey, and assess whether any differences in the quality of the two approaches are attributable to the type of survey, household versus establishment.

2:20 – 3:30
POSTER

DataLumos and Beyond: ICPSR's Tool for Crowdsourced Sharing of Administrative and Other Government Data and Ways to Address its Limitations

- David Bleckley (ICPSR/University of Michigan); Susan Jekielek (ICPSR/University of Michigan); Trent Alexander (ICPSR/University of Michigan)

Increased prioritization of government transparency and accountability along with technological advancements over the past two decades have made government administrative data more widely available than ever before. The long-term accessibility of individual datasets, however, is not always certain. DataLumos is an Inter-university Consortium for Political and Social Research (ICPSR) archive for valuable government data resources. DataLumos uses crowdsourced sharing of publicly-available administrative and other government data to ensure their accessibility now and in the future. The archive has dozens of administrative datasets freely available for public download, and approximately 6000 users have provided, searched for, or accessed data from the website. This presentation discusses the creation of the project, looks back on its first year and a half of implementation, and describes the ways in which administrative data users have benefitted from the tool. Findings from interviews with data users and data intermediaries are presented not only to highlight those benefits but, more importantly, to identify the limitations of DataLumos. The presentation concludes with a discussion of how those limitations have been addressed by ICSPR and ways to address them in the future.

Using Administrative Records in the American Community Survey

- Sandra Clark (U.S. Census Bureau); R. Chase Sawyer (U.S. Census Bureau)

The Census Bureau has made significant progress exploring the use of administrative records in household surveys and the census to accommodate the changing landscape of America's communities and meet the challenges faced by survey researchers. Incorporating administrative records into our data gathering and analysis efforts will have a palpable impact on respondents by reducing the amount of information we request from them. Administrative records may also increase data reliability and provide cost-savings by reducing the need for follow up visits. While there is great potential for the role of administrative records in the future of data collection and processing, there are also challenges to using these data. The Census Bureau recently conducted a test to simulate the use of administrative records to replace responses on the 2015 American Community Survey (ACS) for the questions about property value, property tax, year structure built, and acreage. This presentation discusses the research findings and challenges observed during the test, such as obtaining and securing administrative records, implementing the use of the data into the survey operations, geographic coverage of the administrative records, and the impact the use of the administrative records would have on published ACS estimates.

Using SNAP and TANF Administrative Records and the Transfer Income Model to Evaluate Poverty Measurement

- Kathryn Shantz (U.S. Census Bureau); Liana Fox (U.S. Census Bureau)

Policy leaders today look to quality data and statistics to help inform and guide programmatic decisions. As a result, assessing the quality and validity of major household surveys in capturing accurate program participation is essential. One method for evaluating survey quality is to compare self-reported program participation in surveys to administrative records from the program itself. In this paper, we are interested in understanding two issues. First, how closely do self-reported Supplemental Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF) participation and benefit amounts in the Current Population Survey Annual Social and Economic Supplement (CPS ASEC), as well as SNAP

and TANF participation and benefit amounts corrected for underreporting with the Transfer Income Model, version 3 (TRIM3), align with state-level administrative records. Second, how does replacing values from the CPS ASEC with TRIM3 values or administrative records for SNAP and TANF change poverty measurement in the Supplemental Poverty Measure (SPM).

This abstract was written for submission to the 2018 Administrative Data Research Conference in Washington, DC. It has undergone more limited review than official reports. Do not cite without authors' permission.

The Nonemployer Statistics by Demographics: Leveraging Administrative Data to Meet Stakeholder Needs

- Lisa M. Blumerman (U.S. Census Bureau); Kevin E. Deardorff (U.S. Census Bureau); Aneta Erdie (U.S. Census Bureau); Adela Luque (U.S. Census Bureau)

Like their household counterparts, business surveys endure declining response rates and increasing respondent burden and costs. In this context, administrative records (AR) coupled with AR source expertise and innovative methodologies are playing a critical role in addressing these issues. Here we discuss the creation of a new data product that is based purely on administrative records: the annual Nonemployer Statistics by Demographics or NESD. This product replaces the nonemployer component of a survey with low response rates, and high respondent burden and operational costs. The NESD strives to improve data quality, timeliness and frequency while reducing costs and generating efficiencies in business program management.

Specifically, the Census Bureau has consolidated three business surveys that have traditionally provided U.S. business owner population estimates: the five-year Survey of Business Owners (SBO), the Annual Survey of Entrepreneurs, and Business R&D Survey for Microbusinesses. While employer firm statistics will be available via a new survey (the Annual Business Survey), the NESD represents the continuation of nonemployer demographics estimates previously provided by the SBO. Drawing from existing individual and business AR, the NESD assigns demographics to the universe of approximately 24 million nonemployers, thus creating a new data product that reduces costs and eliminates respondent burden for nonemployers, while improving data timeliness and quality.

The creation of NESD illustrates the value of leveraging AR to create new or replacement business statistics. We discuss our vision for this new product and ways to further develop it to meet the needs of stakeholders and society at large.

Higher Education Administrative Data Elements: Potential for Analysis

- Susan Jekielek (University of Michigan)

Institutions of higher education collect administrative data on students in many forms, including their background (information at application), progress (transcripts), and other formats. In addition, universities collect information at the course, department and university level, such as major/minor requirements, course curricula and materials, and tests, that can be used in combination to provide a picture of student experiences' learning analytics. We will describe a project that is currently being undertaken which has the goals of inventorying data elements and integrating such data across two major Midwest universities.

Linking Administrative Data across Federal Agencies: Outcome Evaluations for NIH Training Programs

- Misty Heggeness (US Census Bureau); Marta Murray Close (US Census Bureau); Andrew Miklos (National Institutes of Health); Nathan Moore (National Institutes of Health)

The National Institute of General Medical Sciences (NIGMS) awards nearly $2.7 billion dollars in federal research and training grants to biomedical researchers across the country to promote basic biomedical sciences and ultimately improve human health. As part of this function, NIGMS provides leadership in training the next generation of biomedical scientists as it supports over a quarter of all research training by the National Institutes of Health. To ensure that it is investing taxpayer money wisely, NIGMS regularly monitors its portfolio including long term outcomes. Assessing the long-term outcomes of federal training programs generally requires a heavy manual curation process. Often, tracking former participants of training programs is a rate-limiting step for performing such evaluations. To address this obstacle, NIGMS partnered with the United States Census Bureau, which serves as the leading source of quality data about the nation's people and economy. This presentation will examine a unique collaboration between NIGMS and the Census Bureau, linking administrative data and lessons learned, and the how data linkages facilitated outcomes evaluations. We will discuss the inception of the collaboration, including the benefits that both federal agencies receive from the partnership, challenges faced in forging a path forward, and some future directions for

the collaboration. The presentation will also demonstrate how this approach could be applied by others to program evaluation, evidence-based policy making, and improved outcomes.

Classifying Families and Households for VA Government Stakeholders using Existing Administrative Records

- Brian J. Goode (Social and Decision Analytics Laboratory - Virginia Tech); Daniel Liden (Social and Decision Analytics Laboratory - Virginia Tech); Jeff Price (Virginia Department of Social Services); Aaron Schroeder (Social and Decision Analytics Division, Biocomplexity Institute of University of Virginia)

The U.S. Census Bureau defines housing units inhabited by individuals related by birth, marriage, or adoption as ""family households."" Housing units inhabited by unrelated individuals are defined as ""nonfamily households."" Accurately classifying types of households can be critical to making informed social service policy decisions. However, making this distinction explicit is not necessarily the current practice at all levels of government. Furthermore, organizations within government may use the definition of family and nonfamily households differently based on individual needs. Therefore, such a distinction may not be uniformly applicable across agencies. Classifying family and nonfamily households using existing administrative records stands to improve on existing methods by: a.) reducing the need for additional resources for collection activities and b.) allowing for tailored definitions to suit inter-agency needs.

Our presentation details the feasibility and development of this capability for stakeholders within the Virginia state government that rely on the distinction between households and families to inform policy decisions. The data sources we consider include the VA Online Automated Services Information System (OASIS) and the Virginia Case Management System (VaCMS). Our methods span the entire data gathering and analysis process from problem definition, data profiling, discovery of additional data sources, and metric creation and validation. We will discuss our findings in this study within the context of identifying families and the considerations needed to account for multiple stakeholders.

Improving Administrative Data Quality at the Initial Stages of the Data Life Cycle

- Megan Davis (Mathematica Policy Research); Emma Kopa (Mathematica Policy Research)

Administrative data have traditionally been collected and employed for uses such as administering and monitoring programs, tracking service and benefit receipt, and meeting regulatory and reporting requirements. Recent technological advances and a focus on less costly and less time-consuming data collection methods have increased the use of administrative data in the public and private sectors for other purposes, such as evaluating program impacts and predictive analytics. The potential for using administrative data effectively for a wide range of purposes is greater if the data are properly collected, well documented, appropriately cleaned and integrated, and accessible to and correctly used by end users, such as program staff and external researchers. Drawing on data governance strategies and our experience collecting and processing administrative data for large-scale evaluations, we will present best practices for improving the quality of administrative data at five key, initial stages of the data life cycle: (1) primary collection, (2) documentation, (3) cleaning, (4) integration, and (5) dissemination. These best practices can improve the reliability of an entity's administrative data and allow end users to experience the many benefits of having high-quality administrative data at their fingertips.

Predicting Spatial Risk of Opioid Overdoses in Providence, RI

- Jordan Butz (University of Pennsylvania); Annie Streetman (University of Pennsylvania)

As the opioid crisis escalates nationally, cities are looking to innovative solutions that might improve the efficacy and accessibility of intervention efforts. As of January 2018, Providence, Rhode Island is implementing a Safe Stations program where people struggling with substance use disorders can come to any of the City's 12 fire stations to be connected with supportive services. This project, produced as part of the University of Pennsylvania Master of Urban Spatial Analytics Spring 2018 Practicum, examines how a predictive modeling approach might assist Providence's Healthy Communities Office in evaluating how well the program's locations are serving Providence residents and where the City could site additional interventions or supplement their

communication efforts. By analyzing administrative health records on opioid overdoses across space, community protective resources, risk factors, and neighborhood characteristics, and harnessing the power of machine learning, this project produced a spatial risk model of opioid overdoses across Providence. The final model was used to create an interactive web application that visualizes risk across the city and allows the City of Providence and other stakeholders to identify high risk areas that are underserved by the current Safe Stations program and which community facilities are best suited to host or advertise additional interventions. An online report provides detailed code and outlines the datasets necessary to replicate a similar process in other cities, allowing the project to serve as a model for cities around the country that are interested in implementing similar initiatives with the greatest impact.

North Carolina Education Research Data Center

- Kara Bonneau **(Duke University)**

In partnership with the North Carolina Department of Public Instruction, the North Carolina Education Research Data Center (NCERDC) at Duke University promotes sound research relevant to education policy by providing access to information about public schools, teachers, and students in North Carolina.

Data available from the NC Education Research Data Center include:

- Individual school size, ethnic composition, statewide accountability results;
- Exceptionality of students;
- Teacher credentials;
- End of Grade and End of Course test results for each student;
- School and student growth and proficiency
- Graduation and school exit
- Advanced Placement coursework
- Grade point average and high school transcripts
- Standardized test scores (SAT, ACT, PSAT)
- Student demographic information; and
- Student infractions leading to short- and long-term suspensions.

With nearly 20 years of experience in managing administrative data and supporting education research across disciplines, the NCERDC is a valuable but underutilized resource that may be of interest to the attendees of this conference.

3:30 – 4:50
B3: RECIPES FOR SUCCESSFUL ACADEMIC-GOVERNMENT COLLABORATIONS

Working with federal government agencies to unlock administrative data

- Jennifer Auer **(Optimal Solutions Group, LLC)**

Federal administrative data is a low-cost and low-burden data source for evidence-based policy making. By linking information from different surveys, or over time, researchers can achieve the sample size and variation needed for advanced econometric methods. However, the personally identifying information (PII) needed to link information means that these data re not available to the public. One solution is to provide technical specifications to the requisite agency(s) to execute the research. This paper outlines the process and pitfalls of drafting specifications for an implementing party who knows more about the data than you do. Drawing on experience from working with the U.S. Census Bureau and knowledge gained from related literatures, such as open-source coding, this paper recommends the depth of description, order of data manipulation and analysis, and requested output to make these collaborative projects successful. A federal administrative data project proposal template is offered. The paper also advises on information that federal agencies can supply to facilitate the use of these important data sources.

Key Factors in the establishment of an academia-government center of public sector administrative data and policy research

- James Ayles **(New Brunswick Department of Health);** Ted McDonald **(New Brunswick Institute for Research, Data and Training)**

A collaborative between the Government of New Brunswick (GNB) and the University of New Brunswick to establish a center of public sector administrative data and policy research was envisioned in 2012. Subsequent work between the parties led to the establishment of the New Brunswick Institute for Research, Data and Training (NB-IRDT) in 2014.

Academia-government partnerships are not unique in Canada, however what sets this apart is: 1) the legislative approach used to support research, 2) scope of administrative data made available, 3) value placed on anonymized linked data, 4) governance overseeing the partnership, and 5) measures taken to ensure the protection of citizens' data.

In 2017, the New Brunswick Act Respecting Research received proclamation. This Act serves to provide clarity and addresses gaps in access and use of personal / health data for research. The Act has opened the doors for NB-IRDT with data owners of public sector organizations. NB-IRDT may now receive pseudonymous personal data from any public sector program collecting personal information.

The partnership is governed by several advisory committees each serving a different role in overseeing the growth of NB-IRDT; overall direction setting being led by a panel of Deputy Ministers and the Clerk (the senior ranking civil servant in GNB.)

The collaboration is well positioned to support public policy research and fosters the use of evidence-based information in the development of government programs and services. The partnership has also helped to encourage new and innovative thinking within GNB about the value of linkable data to support decision-making.

The Washington State Merged Longitudinal Administrative Database

- Jennifer Romich (University of Washington); Mark Long (University of Washington); Scott Allard (University of Washington); Anne Althauser (University of Washington)

This paper describes a uniquely comprehensive database constructed from merged state administrative data. State Unemployment Insurance (UI) systems provide an important source of data for understanding employment effects of policy interventions but have also lack several key types of information: personal demographics, non-earnings income, and household associations. With UI data, researchers can show overall earnings or employment trends or policy impacts, but cannot distinguish whether these trends or impacts differ by race or gender, how they affect families and children, or whether total income or other measure of well-being change. This paper describes a uniquely comprehensive new administrative dataset, the Washington Merged Longitudinal Administrative Database (WMLAD), created by University of Washington researchers to examine distributional and household economic effects of the Seattle $15 minimum wage ordinance, an intervention that more than doubled the federal minimum wage.

WMLAD augments UI data with state administrative voter, licensing, social service, income transfer, and vital statistics records. The union set of all individuals who appear in any of these agency datasets will provide a near-census of state residents and will augment UI records with information on age, sex, race/ethnicity, public assistance receipt, and household membership. In this paper, we describe 1.) our relationship with the Washington State Department of Social and Health Services that permits this data access and allows construction of this dataset using restricted personal identifiers; 2.) the merging and construction process, including imputing race and ethnicity and constructing quasi-households from address co-location; and 3.) planned benchmarking and analysis work.

Integrated Administrative Data for Early Childhood Iowa: A Governance Model to inform Policy and Program Collaboration

- Heather L. Rouse (Iowa State University); Cassandra J. Dorius (Iowa State University); Jeffery Anderson (State of Iowa, Department of Management); Elizabeth JD Richey (State of Iowa, Department of Public Health)

In response to demands on public systems to do more, do better, and cost less, the value of integrated administrative data systems (IDS) for social policy is increasing (Fantuzzo & Culhane, 2016). This is particularly relevant in programming for young children where services are historically fragmented, disconnected from systems serving school-aged children, and siloed among health, human services, and education agencies. Guided by the vision that Iowa's early childhood system will be effectively and efficiently coordinated to support healthy families, we are developing an early childhood IDS to address this disconnection and facilitate relevant and actionable social policy research.

Iowa's IDS is a state-university partnership that acknowledges the need for agencies to retain control of their data while enabling it to be integrated across systems for social policy research. The innovative governance model deliberately incorporates procedures for stakeholder engagement at critical tension points between executive leaders, program managers, researchers, and practitioners. Standing committees (Governance Board, Data Stewardship, and Core team) authorize and implement the work of the IDS, while ad-hoc committees are solicited for specific projects to advise and translate research into practice.

This paper will articulate the Iowa IDS governance model that was informed by means tested principles articulated by the Actionable Intelligence for Social Policy Network. It will include our collaborative development process; articulated mission and principles that guided discussions about legal authorization, governance, and use cases; and the establishment of governance committees to implement our vision for ethical and efficient use of administrative data for social policy.

International Data Access Network (IDAN)

- Roxane Silberman **(CASD-GENES & CNRS Paris, France);** Dana Mueller **(FDZ-IAB Nuremberg, Germany);** Beate Lichtwardt **(UKDS-University of Essex, UK)**

  With legal frameworks changing, administrative data can increasingly be utilised both for official statistics and to facilitate new research, enabling the development of evidence-based policy for the public benefit. Secure access conditions generally apply to using these rich, highly detailed data. However, using data from various sources is difficult when they are fragmented in "silos" between several Research Data Centres (RDCs) as can happen at a national level, and is very likely to be the case at an international level. This is a major obstacle for international comparative research. Based on user consultations, on discussions with international organisations such as OECD and Eurostat and based on lessons learned from projects as, "Data without Boundaries" and the "Nordic Microdata Access Network", IDAN aims to create a concrete operational international framework enabling access to controlled data for research. IDAN, founded in 2018, involves six RDCs from France, Germany, the Netherlands and the United Kingdom. Initially, the partners' access systems are being implemented in each partners' premise based on bilateral agreements. This process involves combining requirements of security and surveillance for Safe Rooms, thus paving the way for next steps toward an integrated RDCs network. This presentation will describe how IDAN is setting up a new concrete environment for researchers to work remotely with data from the other partners within their local RDC. The paper will present first project developments, lessons and impact for research that are also of interest for national contexts where administrative data are held in multiple data centres.

DAY 2 ABSTRACTS

9:00 – 10:20
C1: USES OF ADMINISTRATIVE DATA FOR SUPPORTING EDUCATION POLICY

Long-Term Effects from Early Exposure to Research: Evidence from the NIH "Yellow Berets"

- Pierre Azoulay (MIT & The National Bureau of Economic Research); Wesley Greenblatt (MIT); Misty L. Heggeness (U.S. Census Bureau)

In the late 1960s, the federal government was looking for young, healthy men to enlist in the military to help ensure success in the Vietnam War. Not enough men were voluntarily choosing to enlist. In 1969, the federal government implemented a lottery draft. Recruiters traveled the U.S. encouraging enlistment and explaining the draft requirements. They made visits to medical schools explaining options to newly minted MDs. Those options included: (1) be drafted and (possibly) go to war or (2) enlist in the Public Health Service (PHS) using the skills learned in their medical profession in the U.S. The PHS included an option to travel to Bethesda, Maryland and enlist as a Training Associate (TA) at the National Institutes of Health (NIH) to work in one of the scientific intramural labs on campus and receive training by some of the top medical research scientists in the nation.

For this study, we searched the National Archives for the physical paper applications of those individuals who applied to the NIH Intramural Training Associates program before, during, and after the lottery draft. We digitalized their applications, combed public documents in search of up-to-date career information, and linked them to their publications and patents to-date. We created a rich linked dataset of administrative records from which we examine the impact of early career, high intensity research training on the probability of staying in research and the overall impact on advancing science. This paper describes our results evaluating the impact of this federal program.

Early College High Schools at Scale: Using Administrative Data to Assess the Impacts of an Educational Intervention on Voting and Crime

- Douglas Lauen (University of North Carolina at Chapel Hill), Sarah Fuller (University of North Carolina at Chapel Hill), Tom Swiderski (University of North Carolina at Chapel Hill), Fatih Unlu (Rand Corporation)

Early college high schools (ECHS) are small schools of choice which provide students with the opportunity to earn, at no financial cost to them, two years of transferable college credit or an associate's degree while simultaneously satisfying high school graduation requirements. This promising intervention is aimed at smoothing the transition from high school to college for under-represented minorities and students from economically disadvantaged backgrounds. There are about 80 ECHS in North Carolina, although the model is implemented in many other states as well.

While much is known from prior research about the impacts of the intervention on educational attainment, nothing is known about longer term outcomes such as employment, wages, criminal involvement, and voting behavior. The present study will briefly describe the data collection process, research methods, and preliminary findings on the effects of the intervention on voting and criminal conviction in North Carolina. We will also present results on whether impacts on long term civic outcomes are mediated by educational attainment. Quasi-experimental impacts have been validated against impacts generated from a randomized controlled trial of the same intervention in a subset of the sites during the same time period.

The team assembled personally-identified population level statewide administrative data on all NC high school students (including ECHS) and linked it to records housed at community colleges, universities, the Department of Public Safety (incarceration), and Board of Elections (voting). Together this effort comprises one of the more comprehensive administrative data collection efforts linking student level K-12, postsecondary, and longer-term outcomes.

Demonstrating the Use of Linked Data to Advance Education Research

- Sonya R. Porter (US Census Bureau); Nick Pharris-Ciurej (US Census Bureau); Emily Penner (University of California, Irvine); Quentin Brummet (NORC)

  Linking K-12 data on students and teachers to Internal Revenue Service (IRS) information allows us to answer questions that are difficult to answer using survey data or educational administrative data alone. We describe two research projects that demonstrate the importance of using linked administrative data to further research on education and inform policy discussions. In the first research project, using linked IRS income tax data to school administrative records for all 8th graders in one California public school district and all K-12th graders in Oregon public schools, we examine how well free and reduced price lunch (FRPL) enrollment captures student disadvantage. We find that FRPL categories capture relatively little variation in household income. However, FRPL captures elements of educational disadvantage that IRS-reported household income data do not. In the second research project, using data on teachers from a large California school district linked to IRS records and the Business Register, we examine what teachers do after they leave teaching. Preliminary findings suggest that many teachers leave the workforce after they leave teaching. Teachers that continue to work after leaving our school district often do so in a nearby school district, and often see a modest increase in their earnings in their new positions.

Returns to Late Aged College Degrees

- Sally Wallace (Georgia State University); Thomas Mroz (Georgia State University); Alex Hathaway (Georgia State University)

  The benefits of a college education are well documented. However, the majority of existing research focuses on students who matriculate soon after high school graduation. There is little empirical evidence illustrating whether a college degree is similarly beneficial to those already in the workforce, particularly individuals over 50. Nonetheless, the coming years will see the dramatic growth of older individuals, many of whom will continue to be active in the labor force, and policymakers would benefit from effective strategies to improve the labor market outcomes of older individuals.

  This research proposes to evaluate the labor market outcomes of individuals in Georgia who obtain a bachelor's degree at age 50 or older by merging state-level individual level labor force (Dpt of Labor) with individual level educational data from the University System of Georgia (USG). Specifically, we explore whether these later-age degrees result in employment opportunities with higher wages and increased retention in the labor force beyond the traditional retirement age of 65 than those who do not attain a bachelor's degree. The results will provide policymakers across the United States with information to make informed decisions regarding higher education incentives and policies for older students.

9:00 – 10:20
D1: USES OF ADMINISTRATIVE DATA FOR SUPPORTING PUBLIC PROGRAMS AND PUBLIC HEALTH POLICY

Evaluating the impact of workers' compensation policy in Australia using insurance claims data and comparative quasi-experimental methods

- Alex Collie (Monash University, Australia), Tyler Lane (Monash University, Australia), Shannon Gray (Monash University, Australia)

  Australia, like the USA, has state-based workers' compensation (WC) systems that provide income support, healthcare and rehabilitation for injured and ill workers. The eleven major Australian WC systems provide coverage for over 90% of the labor force and accept approximately one quarter of a million new claims per annum. Governments commonly use changes in scheme design (most often enacted through legislative amendment) to influence WC system performance including rates of claiming, costs and return to work (RTW) outcomes. Using a national, longitudinal, case level dataset of WC insurance claims data, we evaluated the impact of multiple, state level legislative amendments. The impact of legislative amendments in the states of South Australia (year of 2009), Tasmania (2010), Victoria (2010) and New South Wales (2012) were evaluated using interrupted time series analysis. Outcomes included volume and incidence of accepted WC claims, employer and insurer claim processing timeframes, and duration of work disability. Major findings include (1) the Tasmanian amendments designed to improve RTW outcomes failed; (2) the South Australian amendments designed to encourage early

employer claim lodgment were partially effective; (3) the New South Wales amendments designed to ensure the financial viability of the WC scheme reduced access to benefits and disproportionately affected workers with occupational disease and mental health conditions; (4) the Victorian amendments designed to increase benefit generosity led to an increase in claims and longer duration of disability. Study findings demonstrate both intended and unintended consequences of WC system reform, and provide an evidence base for future reform.

Fusing Administrative Data to Combat the Opioid Crisis

- Kevin P. Conway **(RTI International);** Camille Gourdet **(RTI International)**

Opioid-related overdose deaths remain the leading cause of unintentional injury fatalities in the United States. State lawmakers have responded to this crisis by establishing a regulatory environment that extends various legal protections to persons who may help save the life of someone experiencing an opioid-related overdose. Most states now protect specific parties (e.g., doctors, pharmacists, first responders, laypersons) from civil or criminal liability who prescribe, dispense, possess or administer an opioid antagonist in accordance with the provisions of the state's law. In addition to standing orders that facilitate access to opioid antagonists, many states offer legal protection to "Good Samaritans" seeking medical and emergency assistance for a person experiencing an overdose. Some states additionally mandate that addiction-treatment services be offered in conjunction with the dispensing of an opioid antagonist, whereas others designate revenue to purchase opiate antagonists or to fund treatment programs.

Little is known about the potential impact of such regulatory actions on the opioid crisis. RTI's Data Fusion Center seeks to meet this need by combining administrative data across sources and systems to inform research and policy. The current paper describes the Data Fusion Center and presents preliminary results from a study that predicts opioid-related overdose deaths based on the existence and strength of opioid-related state laws among 50 states from 2006 to 2016. Policy data were webscraped from state agencies, systematically coded, and associated with target outcomes sourced from CDC. Study findings may help inform lawmakers and stakeholders in prioritizing data-driven policy responses to the opioid crisis.

Developing a Data-driven Approach to inform Planning in County Health and Human Services Departments in the Context of a Case Study on Obesity

- Ian Crandell **(Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech);** Joshua Goldstein **(Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech);** Vicki Lancaster **(Social and Decision Analytics Division, Biocomplexity Institute of University of Virginia);** Dutton Sophia **(Fairfax County, Office of Strategy Management for Human Services)**

Since the 1970s, the obesity rate has steadily increased due to growing availability of food and declining physical activity. The existing environments within a community, including active recreation opportunities, access to healthy food options, the built environment, and transportation options, can moderate obesity. In Virginia, Fairfax County Health and Human Services (HHS) system is interested in developing the capacity for data-driven approaches to gain insights on current and future issues, such as obesity, to characterize factors at the county and sub-county level, and to use these insights to inform policy options. In exploring these questions, we developed statistical methods to combined data from a multitude of different sources including local administrative data (e.g., tax assessments, land use, student surveys), place-based data, and federal collections. Using synthetic data methods based on imputation, we recomputed American Community Survey statistics for non-Census tract geographic regions for political districts and high school attendance areas. We combined this with environmental factors, such as land dedicated to parks and recreation facilities, as well as measures of the density of healthy and unhealthy food locations to create a map of potentially obesogenic factors. Finally, we combined these data sources with Fairfax County's youth survey and trained a random forest model to predict the effects of the environment on healthy food consumption and exercise. Our analysis highlights the need for (administrative) data at a fine scale and recommends policy changes concerning the recording and sharing of local data to better inform the policy and program development.

The Use of Blended Data to Improve Public Assistance Programs: Results from a Partnership between the U.S. Census Bureau, USDA, and State Program Agencies

Benjamin Cerf (U.S. Census Bureau); Thomas B. Foster (U.S. Census Bureau); Mark Leach (U.S. Census Bureau); Rachel Shattuck (U.S. Census Bureau)

The Census Bureau is partnering with state public assistance agencies to acquire program participation data and estimate new statistics that deepen a state's understanding of program participants and improve outreach efforts to those who are eligible but do not participate. In collaboration with the Economic Research Service and the Food and Nutrition Service within the United States Department of Agriculture, the Census Bureau obtains individual-level program participation administrative records (AR) data for three state programs, the Supplemental Nutrition Assistance Program (SNAP), Temporary Aid for Needy Families (TANF), and the Special Supplemental Nutrition Program for Women, Infants and Children (WIC). The Census Bureau constructs a unique data set for each state program by linking the AR data to survey response data for the same individuals. These linked data enable the Census Bureau to model which survey respondents are eligible for program participation and also to observe which eligible individuals participate in the program. The Census Bureau then estimates eligibility and participation rates by a variety of demographic and economic characteristics and by county. The individual-level data also enable the Census Bureau to construct a statistical profile of eligible individuals and families that do not participate to assist state agencies with their outreach programs. All statistical results provided back to state agencies in table reports and data visualizations are reviewed to insure that individual identities are protected and not disclosed. This paper will present results for several state programs that have partnered the Census Bureau.

The power of linked data: Evaluating diverse multi-program projects designed to reduce welfare dependence

Jenny Povey (The University of Queensland); Janeen Baxter (The University of Queensland); Christopher Ambrey (The University of Queensland); Guyonne Kalb (The University of Melbourne); David Ribar (The University of Melbourne); Mark Western (The University of Queensland)

This presentation showcases the innovative use of linked government administrative data in Australia to evaluate a range of diverse social interventions aimed at supporting vulnerable groups to achieve economic independence. The interventions were developed and funded as part of the Australian Priority Investment Approach to Welfare, an approach supported by actuarial analyses of administrative data designed to deliver targeted support for groups at-risk of long-term welfare dependence. In 2018, the Australian Government, commissioned an impact evaluation to assess the effectiveness of the approach in achieving its intended outcomes. The evaluation is based on analyses of linked administrative data to assess the extent to which the new interventions enabled pathways out of welfare dependence. Our presentation will outline the strengths and weaknesses of using government administrative data to evaluate the outcomes. Strengths include easy comparison across diverse programs designed to achieve the same goals; reduced respondent reporting burdens; robust quasi-experimental techniques such as a matching design based on exact matching on a few key characteristics and/or propensity score matching on a broad range of pre-program characteristics; and evidence-based investment practice decisions. Weaknesses include the adoption of an observational rather than experimental design and the lack of information on some social characteristics such as orientations to work, attitudes and social values. The presentation not only assesses the compilation of administrative data used for the first time to evaluate multi-program projects, it will also describe how these data feed into visual interactive dashboards used to monitor the outcomes of the interventions.

DAY 2 ABSTRACTS

10:40 – 12:00
C2: USES OF ADMINISTRATIVE DATA FOR SUPPORTING CHILD WELFARE POLICY

Using Integrated Data to Design and Support Pay For Success Interventions

- Claudia Coulton (Case Western Reserve University); Meghan Salas Atwell (Case Western Reserve University); Francisca Richter (Case Western Reserve University); Elizabeth Anthony (Case Western Reserve University)

Pay for Success (PFS) interventions are increasingly being implemented in the U.S. and worldwide to assess social programs under a risk-sharing financial agreement between the public and private sectors. They seek to mitigate risk for the public sector and promote wider experimentation of programs to improve social outcomes. PFS contracts encourage coordination and alignment of goals, outcomes, and metrics across all agents involved - government, service providers, service recipients, funders and investors. Accordingly, these interventions rely heavily on access to high quality data and analysis, making integrated data systems (IDS) valuable assets to support the design, implementation, and evaluation phases of these projects.

The ChildHood Integrated Longitudinal Data (CHILD) System, one of the most comprehensive county-level IDS in the nation, has been used to support and inform two Pay for Success projects in Cuyahoga County (Cleveland). Partnering for Family Success is a county-level intervention in the areas of child welfare and housing instability, now into its fourth year of operation. While the intervention was implemented under a randomized controlled trial, analysis with CHILD proved instrumental to inform the project design and address challenges in program implementation. CHILD has also been used to study the feasibility of PFS as a model to expand high quality preschool, under a grant awarded to eight communities nationally. A case study of both initiatives will be presented, highlighting the role of integrated data in supporting and facilitating PFS design and analysis of outcomes, challenges encountered and lessons learned.

Policy and resource optimization based on spatial risk prediction models for child maltreatment

- Dyann Daley (Predict-Align-Prevent, Inc.)

Spatial risk prediction models describe the features of small spatial units which support or attract child maltreatment behaviors. Due to shared risk factors, community problems such as infant mortality, poor educational readiness, injury-related deaths, and a host of physical and mental health problems associated with toxic stress co-occur in these small areas. Based on administrative data, this spatial intelligence brings cross-sector stakeholders together to collaboratively plan for optimization of critical supports, capacity development for vital services, improvement of professional response, development of supportive infrastructure, and creation of healthier social norms. We present the policy and resource optimization strategies being developed in locations implementing the Predict-Align-Prevent program for child maltreatment prevention.

Effect of Psychotropic Medication on Foster Care Experience and Outcomes: A Causal Analysis using Administrative Data

- Angela You Gwaltney (University of North Carolina, Chapel Hill)

Children in foster care experienced abuse, neglect, or dependency, and for the safety and well-being of the child, must be taken out of their biological home. Not surprisingly, children in foster care have higher rates of serious emotional and behavioral problems. Although pharmacological treatments can be an important component of the treatment plan, there seems to be a higher rate of use than would be expected. An estimated 13-25% of foster children are prescribed mind- and mood-altering medication vs. 4% in the general population.

Children in foster care are considered a vulnerable population and research involving these children justifiably requires additional measures to ensure their protection. As a result, studies on the use of psychotropic medication among youth in foster care have relied primarily on secondary data, typically administrative data. This study

used linked administrative datasets to rigorously examine the effect of psychotropic medication on foster care experiences and outcomes among children who entered foster care in North Carolina between March 2006 and June 2012. The dataset was constructed by linking the North Carolina's child welfare administrative records (also known as the Services Information System [SIS]) with the Medicaid claims database (also known as the Eligibility Information System [EIS]) for medical and mental health services received by the foster youth. Inverse probability of treatment weighting was calculated and applied to mimic a randomized study. Results revealed that children on medication stayed in care longer, less likely to experience placement disruption, and more likely to exit to adoption.

A Typology of Transition-Age Youth

- Jessica Raithel **(Center for Innovation through Data Intelligence);** Andrew Wallace **(Center for Innovation through Data Intelligence);** Maryanne Schretzman **(Center for Innovation through Data Intelligence);** Eileen Johns **(Center for Innovation through Data Intelligence)**

  Young adulthood is a time of transition which poses particular challenges for youth who are homeless or at risk of homelessness, including those exiting foster care. The instability of being homeless puts youth at greater risk of many poor outcomes. Connection to relevant housing resources and services are critical to ensure that young adults have the opportunity to succeed. Better aligning youths' needs with relevant housing resources can help young adults become and remain stably housed, leading to better lifetime outcomes. This study presents a typology of young adults who exit foster care and residential programs for homeless young adults, including emergency shelters and transitional living programs. The study uses administrative data to follow a cohort of 8,795 young adults, including young parents and unaccompanied young adults from ages 18 through 21, who exited foster care or homeless services. Using sequence analysis, subsequent service use after exit, including utilization of homeless services, hospitals, jail, subsidized housing, and supportive housing, was used to build three-year trajectories of service use patterns of youth. These patterns were then grouped together based on similarity using cluster analysis to form six distinct groups of youth: (1) Minimal Service Use, (2) Later Homeless Experience, (3) Earlier Homeless Experience, (4) Consistent Subsidized Housing, (5) Consistent Supportive Housing, and (6) Frequent Jail Stays. Profiles were developed for each typology to comprehensively, but concisely, describe differences in the characteristics of each group of youth. Models were also developed to determine factors that were predictive of each typology. This typology is being used to inform prioritization processes for housing resources and to better understand how to target programs based on potential pathways of youth.

Developing a spatial risk prediction model for child maltreatment

- Ken Steif **(University of Pennsylvania);** Matthew Harris **(Urban Spatial);** Dyann Daley **(Predict-Align-Prevent)**

  While predicting child maltreatment risk at the household level is useful for allocating limited child welfare resources, significant privacy, data integration, data governance and legal hurdles make such an algorithm economically and politically difficult to put into production. In this project, we take a different approach to child maltreatment risk prediction, developing machine learning models that predict, not for a household but for a small spatial areal unit, such as the block. The only private health data required for this use case are geocoded maltreatment events. We present the results of a machine learning analysis in Richmond Virginia, including exploratory analysis, feature engineering, model development and validation. We then interpret our models in a resource allocation context.

10:40 – 12:00
D2: OTHER USES OF ADMINISTRATIVE DATA TO IMPROVE DECISION-MAKING AND PROGRAM EVALUATION

Stretching a Buck: Using Administrative Data to Inform Continuous Quality Improvement

- Sadaf Asrar **(Optimal Solutions Group, LLC)**

  While the nFORM administrative data system is used to collect operations and performance data from HHS funded Healthy Marriage and Responsible Fatherhood (HMRF) grantees to evaluate program performance using top line measures like enrollment and attendance of workshops, the rich data collected through the system provides an unparalleled window into the workings of the HMRF programs and the population they serve. This paper describes how raw data exported from the nFORM system have been used to develop econometric models to

understand the relationship of demographic and socio-economic characteristics on program enrollment and attendance, the effectiveness of incentives and behavioral nudges on program participation, as well as changes in behaviors and attitudes due to the intervention. Moreover, the paper discusses how patterns revealed through mining the raw nFORM data combined with other administrative data has provided insights into deficiencies in outreach and recruitment efforts, and highlights how the relative effectiveness of steps taken to remedy the deficiencies can also be tracked using the data. Lastly, the paper presents recommendations and best practices in using nFORM and other similar administrative data to inform continuous quality improvement of HMRF programs without stretching the budget.

Leveraging U.S. Army Administrative Data for Individual and Team Performance

- Joshua Goldstein **(Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech);** Vicki Lancaster **(Social and Decision Analytics Division, Biocomplexity Institute of University of Virginia);** Nathaniel J. Ratcliff **(U.S. Army Research Institute for the Behavioral and Social Sciences);** Joel A. Thurston **(U.S. Army Research Institute for the Behavioral and Social Sciences)**

The Army possesses vast amounts of administrative (archival) data about Soldiers. These data sources include screening tests, personnel action codes, training scores, global assessments, physical fitness scores, and more. However, the Army has yet to integrate these data to create a holistic operating picture. Our research focuses on repurposing Army administrative data to (1) operationalize social constructs of interest to the Army (e.g., Army Values, Warrior Ethos) and (2) model the predictive relationship between these constructs and individual (i.e., Soldier) and team (i.e., unit) performance and readiness. The goal of the project is to provide people analytics models to Army leadership for the purposes of optimizing human capital management decisions.

Our talk will describe the theoretical underpinnings of our human performance model, drawing on disciplines such as social and industrial/organizational psychology, as well as our experience gaining access to and working with Army administrative data sources. Access to the archival administrative data is provided through the Army Analytics Group (AAG), Person-event Data Environment (PDE). The PDE is a business intelligence platform that has two central functions: (1) to provide a secure repository for data sources on U.S. military personnel; and (2) to provide a secure collaborative work environment where researchers can access unclassified but sensitive military data.

Aboriginal life pathways through multiple human service domains; administrative data linkage for policy

- Francis Mitrou **(The University of Western Australia);** Stephen Zubrick **(The University of Western Australia);** Glenn Pearson **(Telethon Kids Institute);** Anna Ferrante **(Curtin University);** Melissa O'Donnell **(The University of Western Australia);** Sarah Johnson **(Telethon Kids Institute);** Helen Milroy **(The University of Western Australia);** Ngiare Brown **(South Australian Health and Medical Research Institute);** Jonathan Carapetis **(The University of Western Australia)**

Aboriginal children and families face the highest levels of disadvantage of any population group in Australia across health, education, child protection, justice and other human service domains, but longitudinal data to inform policy is scant. The Western Australian Aboriginal Child Health Survey (WAACHS) is a population representative cross-sectional child development study of over 5,000 randomly selected children aged 0-17 years, plus their families and schools, conducted between 2000 and 2002. This project seeks to leverage the WAACHS by linking the survey data for all participants with State administrative human services data registers from the previous 30+ years, to develop a major program of work in Aboriginal Human Development that would be unique in the world. This presentation describes the project history, novel survey linkage methodology, and project aims in the policy domain.

Trust Central Eases Funding Decisions With Data

- Marden F. Munoz **(The Children's Trust);** Stephanie Sylvestre **(The Children's Trust)**

The Children's Trust is a local government taxing authority that uses property tax dollars to fund programming for Miami-Dade County children and families. To become more efficient, Trust Central was created and now automates our full business cycle. Data flow is solicitation > contracting > program metrics > solicitation. Agencies apply for funding, contract to provide services, report their progress. Their progress is used to determine future solicitation criteria. Data automatically flows from one module to another.

Trust Central allowed us to move from using 5 data points to make funding decisions to 24 data points. We were able to look across our various initiatives to ensure that our funding decisions were equitable. Funding decisions were backed by data and easy to share with applicants. We created context and communicated funding decisions in a way that reduced emotional conflicts and appeals. As a reference point, we had 96 appeal meetings last funding cycle - this funding cycle no appeal meetings and only 19 review meetings; a cost savings of $15,850 in meetings. Another reference point, we spent 1 week reviewing data to make funding decisions last funding cycle - this funding cycle we spent 4 weeks reviewing data in a more meaningful and valuable way. We made $68M in funding decisions without any negative feedback from the community. Our relationship with the community pivoted from negative to positive. This is a first for us! We are now positioned to be a mentor for both governmental and non-governmental funders.

Use and Application of Federal Advisory Committee Act (FACA) Database

- Natsuko Nicholls **(Institute for Research on Innovation and Science (IRIS) / University of Michigan Institute for Social Research);** Jason Owen-Smith **(Institute for Research on Innovation and Science (IRIS) / University of Michigan Institute for Social Research)**

University experts can offer uniquely valuable insights for informing policy based on expertise they develop through research. The application of knowledge through public service is an important and understudied mechanism for translating academic expertise to government and other communities. Today universities encourage researchers to engage in public service, and often they actively provide institutional support to create a culture and environment where such pro bono work is regarded as an important activity by the research community. Yet the question remains as to whether or not a systematic mechanism exists to track, record, and measure the value of university expertise influencing policy within the context of research. We explore a useful but underutilized administrative data source, the Federal Advisory Committee Act (FACA) database, with an eye towards linking the federal service data to other sources in order to measure research impact in a sociopolitical setting. This publicly available dataset contains rich information on federal advisory committees that play an important role in shaping national programs and policies. Each year an average of 900 advisory committees with more than 60,000 members have provided either policy or grant review advice in 40 different issue areas. Our exploratory findings suggest a steady increase of academics in federal service, the different level of federal service contribution by universities, and the association between federal service and university R&D spending. We also discuss the importance of data cleaning when using administrative data for research and data linkage methods when linking federal service data to university research spending records.

1:40-2:30
RAPID FIRE TALKS

Reproducible research in administrative data center

- Lars Vilhuber **(Cornell University);** Carl Lagoze **(University of Michigan)**

The recent concern about the reproducibility of research results has not yet been robustly incorporated into methods of providing and accessing administrative data, casting doubts on the validity of research based on such data. Reproducibility depends on disaggregating and exposing the multiple components of the research - data, software, workflows, and provenance - to other researchers and providing adequate metadata to make these components usable. The key worry is access: the authors of a study that uses administrative data often cannot themselves deposit the data with the journal, thereby impairing easy access to those data and consequently impeding reproducibility. This suggests a critical role for administrative data centers. We argue, that data held by ADRF do have attributes that lend themselves to reproducibility exercises, though this may, at present, not always be communicated correctly. We describe how ADRF can and should promote reproducibility through a number of components.

DAY 2 ABSTRACTS

Is Limited English Proficiency a Barrier to Homeownership?

- Edward Golding (Urban Institute); Laurie Goodman (Urban Institute); Sarah Strochak (Urban Institute)

Nearly 5.3 million US heads of household have limited or no ability to speak English. The connections between race or ethnicity and homeownership have been documented, but there has been little work to explain the relationship between the ability to speak English and homeownership. As homeownership is a primary tool for wealth building and financial stability, it is useful to understand the challenges this population faces in accessing homeownership.

This brief first defines and identifies the limited English proficient (LEP) population in the United States. Using descriptive analysis and regression models, we find that at the zip code level, higher rates of limited English proficiency are associated with lower homeownership rates. If we control for other factors that influence homeownership (e.g., income, age, and race or ethnicity), zip codes with the highest concentrations of LEP residents have homeownership rates 5 percentage points lower than zip codes with the median concentration of LEP residents. In other words, limited English proficiency is a barrier to homeownership.

Complying with New Address Requirements for the 2020 Census: Using Open Administrative Records and Delivery Point Validation Data to Discover Verified Multi-Family Unit Addresses in Arlington County, Virginia

- Aaron D. Schroeder (Social and Decision Analytics Division, Biocomplexity Institute of University of Virginia)

This presentation will provide an overview and demonstration of how the Social Decision Analytics Lab at Virginia Tech (SDAL) employed the use of multiple publicly available real estate administrative datasets, a novel approach to unit number estimation, and the use of a USPS Delivery Point Verification service to help Arlington, Virginia successfully comply with a new multi-family unit address verification requirement of the 2020 Census.

In 1994, Congress established the Local Update of Census Addresses Program (LUCA) which provides the only opportunity for local jurisdictions to review and provide corrections to the U.S. Census Bureau's residential address list. While initial LUCA requirements only required the provision of the street address for both multi-family and single-family units, for the 2020 Census, the requirements have been expanded to include provision of address information for each unit within each apartment or condominium building (i.e. unit numbers and/or letters).

Most localities in Virginia do not regularly collect address data for every multi-family unit. This circumstance is also not limited to Virginia as can be seen in a recent New York Times article discussing the same issue being faced by New York City.

While Arlington chose to conduct an online survey to capture these addresses, given historic survey response rates, it was estimated that 2-3 weeks of full-time staff hours would be required to follow-up with individual building managers. To significantly reduce these estimated staff hours, Arlington partnered with SDAL to develop an automated and replicable approach to multi-family unit address discovery and verification.

Blended data: a novel opportunity to advance survey operations and knowledge of the US economy and population

- Nikolas Pharris-Ciurej (US Census Bureau); Lisa M. Blumerman (US Census Bureau); Mark A. Leach (US Census Bureau)

This presentation highlights the Census Bureau's Center for Administrative Records Research and Applications usage of blended data to improve survey design, implementation, and quality and to create products that provide information on the US economy and population. Related to improvements in survey frame development we discuss two applications: the implementation of administrative records and third-party data in the development of the National Teacher and Principal Survey frame and in current research on the National Survey of College Graduates (NSCG) frame. Then, we broadly discuss ongoing research related to efforts to reduce respondent burden through the use of administrative records and the removal of survey questions. Specifically, we present findings from research that evaluates the coverage and quality of administrative record and third-party data to determine its suitability for item replacement or supplementation of household-level responses to the American Community Survey and individual-level responses to questions on the NSCG. These findings exemplify the promise of administrative records for item replacement or supplementation as well as the complexity related to integrating administrative records into a sample survey production process. Finally, we focus on the creation of new data products using

linked administrative records and survey data. Specifically, we showcase an interactive data visualization tool that displays SNAP, WIC, and TANF program participation by geographic location and creates profiles of program participants and eligible non-participants to aid in program outreach and in service delivery to current participants.

Using Administrative Data to Reduce Respondent Burden in Facility Data Collection

- Andrea Mayfield **(NORC at the University of Chicago);** Rachel Carnahan **(NORC at the University of Chicago);** Felicia LeClere **(NORC at the University of Chicago)**

The Medicare Current Beneficiary Survey (MCBS) is a longitudinal panel, multi-purpose survey of a nationally representative sample of the Medicare population, conducted by the Centers for Medicare & Medicaid Services (CMS) through a contract with NORC at the University of Chicago. The MCBS collects detailed data about Medicare beneficiaries, including health care use and access barriers, health care expenditures, and factors that affect health care utilization. Data are collected via in-person interviews for both community and facility dwelling beneficiaries. The Facility instrument, which is administered to facility staff, was designed to incorporate many items that are currently reported to CMS for Medicare-certified facilities through the Long-Term Care Minimum Data Set (MDS) and the Certification and Survey Provider Enhanced Reports (CASPER) process. Facility data collection is not restricted to persons who are in Medicare-certified facilities but the data collection instrument is similar for both. This paper focuses on efforts to incorporate data from the appropriate MDS and CASPER administrative records in lieu of interview data for those beneficiaries who are located in Medicare-certified facilities. We will describe the analysis necessary to test the feasibility of using administrative records for selective beneficiaries, while maintaining comparability with survey data collected for those beneficiaries in other facility types. We will also describe how the analysis informed a redesign of the MCBS Facility instrument and data file integration process and how these changes reduce burden for interviewers and facility staff.

The Role of the Bundesbank Microdata Production in Times of Big Data: The Need for Data Access, Data Sharing and for an Integrated Digital Information System

- Christian Hirsch **(Deutsche Bundesbank)**

Deutsche Bundesbank - as other central banks - collects monetary, financial and external sector statistical data, comprehensive sets of indicators and seasonally adjusted business statistics. So, the Bundesbank is one of the largest data producers in Germany and its data are of high quality. This applies also to its micro data - quality-tested administrative data covering the fields of banks, securities, enterprises and household finance.

To meet the demand of data users and data compilers for (granular) data sharing and to facilitate the implementation of the G20-Recommendation II.20 of DGI-2, the Bundesbank provides free of charge access for external independent researchers to its (linked) microdata for research purposes in its Research Data and Service Centre (RDSC).

To improve the knowledge RDSC together with NYU (Julia Lane) are developing an Integrated Digital Information System (IDIS). IDIS is a dynamic and adaptive repository which connects data producers, RDSC and Bundesbank researchers, by building a community around research projects, data sets and publications (knowledge map) The knowledge map turns fragmented knowledge produced at all stages of the research process into discoverable and reusable knowledge. Second, by incorporating possibilities for all data users to feed back their information, the knowledge map becomes dynamic. So, IDIS creates value by making discovery of data and related projects, people, and publications at Bundesbank more comprehensive and efficient through storage of usable knowledge in a repository. It also enables analysis about research using modern statistical tools.

Program Evaluation Using Multiple Datasets

- Brittany Borg **(US Small Business Administration);** Terell Lasane **(US Small Business Administration)**

The US Small Business Administration conducts program evaluations to determine if programs are functioning properly, producing the desired outcomes, and to identify areas for improvement. The 8(a) Business Development Program is a business assistance program for small

disadvantaged businesses. 8(a) certified firms are able to obtain special contracting opportunities through the federal government. SBA's 8(a) Program offers a broad scope of assistance to firms, including specialized training through SBA's 7(j) training program. This program evaluation links several datasets (8(a) certification data, 7(j) training data, federal government contracting data, and federal government contractor performance data) to determine if 7(j) training affects the contracting outcomes of 8(a) certified firms. No new data is collected or utilized beyond previously collected administrative data.

Overcoming policy barriers to administrative data sharing through an inclusive civil society coalition

- Michael Lenczner (Powered by Data); Jonathan McPhedran Waitzer (Powered by Data)

In the Canadian context, there is no coordinated policy agenda for increasing social impact through administrative data use. The nonprofit sector is uniquely positioned to advocate for a strong political commitment to linked administrative data. As a sector, it could directly benefit from that data for impact evaluation and for advocacy. It is also closest to the people who are most likely to be negatively impacted by the resulting surveillance and stigmatization.

We are building a network of social service organizations, foundations, and advocacy groups to explore the possibility of creating a shared policy agenda. We have developed a coalition model that engages these unequally resourced stakeholders on equal footing - with the goal of enabling fully-informed and equitable participation.

This coalition is working to develop a set of conditions for increased administrative data linking that reflect the shared interests of funders, service providers, advocacy groups, and beneficiary communities. We are also researching the legislative and policy changes required to enable that desired outcome.

Beyond developing and advancing a shared policy agenda, this initiative also aims to deliver long-term outcomes involving increased data policy literacy among Canadian nonprofits. The coalition itself represents collaborative infrastructure to enable ongoing, coordinated input from the nonprofit sector on key questions of data governance and policy.

This equity-focused, multi-stakeholder coalition approach to digital policy development represents a significant innovation in public engagement. We are excited to discuss our approach, initial results, and key learnings with conference participants.

Immigrant Careers and Networks

- David Card (University of California at Berkeley); Benoit Dostie (HEC Montréal); Bery Li (Statistics Canada); Daniel Parent (HEC Montréal)

The process of economic assimilation by immigrants and how their career and labour market earnings trajectory evolve over time is of obvious concern to policy makers. Knowing how the economic status of immigrants change over time following their arrival and as they integrate the labour market would inform us on the specific channels leading to improved outcomes. One would also like to know whether, and to what extent, the career dynamics of immigrants differ from those of natives. In this paper, we look at how immigrants sort into different firms as they assimilate following arrival and how job referral networks play a role in that process. To do so, we use the Canadian employer employee dynamic database (a link between the worker and the firm's tax files), matched to the Longitudinal Immigration Database (IMDB) to estimate the growth in earnings with time in Canada (basically the rise in occupational status) as well as changes in the characteristics of the jobs, such as the fraction of native Canadians at the firm and the size of the firm. To investigate how networking effects may affect immigrant outcomes and what fraction of the firm effects are due to social networks, we adapt and generalize the strategy used in Schmutte (2015) to relate the estimated firm effects to the social networks connecting workers to jobs in high-paying firms.