“Building Evidence by Linkage to Data at the U.S. Census Bureau” Discussant Comments

Ben Bolitzer, Department of Commerce

October 25, 2023

Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Department of Commerce

Linking Improves Administrative Data

"The information collected in administrative records is determined by the entity administering the program. In probability surveys, data collection is tailored to information needed to calculate statistics." – NAS, 2023

Linked data addresses some of limitations of strictly administrative data:

- Lack of data not essential to administrating the program – demographics, etc.
- If the data is on program participants – what is the counterfactual?

My Discussant Comments

- Overview of the papers
- Describe the ways each paper leverages linking to build better evidence than administrative data alone.
- Offer suggestions for future work – things to consider – things that I wondered about

The Labor Market Returns to Earning Industry Credentials

- Goal: Estimate the labor market returns (employment, earnings) from an industry credential
- Background: Paper cites literature on both the growing prevalence of industry credentials, and the challenges of measuring industry credentials in existing "standard" data sources, although additional questions on credentials were added to the CPS in 2017
- Methods
  - Graphical Analysis: Comparing the time series before and after receiving degree – stratified by educational attainment, and age at last credential
  - Regression Modeling:
    - (Matching) Construct a control sample with ACS of those not in the sample and without a college degree.
    - (Regression) Include as co-variates variables included in the matching process

The Labor Market Returns to Earning Industry Credentials

- Data: Paper links administrative data from a collaboration between NAM and NSC.
  - Industry credentials from four organizations were linked with information on education attainment from NSC.
  - Approximately 54 percent of the credentials were linked to an ID, of those about 85 percent were linked to survey (ACS), W2 and 1040 data.
- Findings:
  - In both the graphical and regression analyses, they find that adding a credential increases earnings, and employment. The results are comparable between the methods – about $5,000 for earnings, and 13 percent for employment, depending on the group.

Paper 1: Demonstration of the Value of Linked Data

Lack of data not essential to administrating the program

- Linking with the ACS data allowed to see the demographics of the program participants. Without this linking – the authors would not have been able to even do the graphical analysis.

If the data is on program participants – what is the counterfactual?

- Survey data allowed the authors to develop the control group, based on ACS respondents. This was much stronger than if they had relied on simply the program data.

Paper 1: Things I wondered about

- As the authors note, a limitation is that it's unknown whether in the ACS data the survey participants have credentials.
- I think that a gap to understanding the external validity of the paper is that we are not given much information about the credential process itself. Are these programs "standard"
- The authors noted the "ashenfelter dip", but it seems like there was a downward trend well before the three years. Was this dip examined?
- Whether it made sense to match on industry. It seems to me that part of the argument for credentials is that it would allow individuals to switch industries. By controlling for industry, are you understating the effect of the credential? On the other hand, you also want to match the distributions of the participants to the source industries to make them comparable.
- Even after the matching, there still was a gap in pre-participation earnings. As noted in the paper, does this imply that there is still unobserved heterogeneity?
- Standard errors -> it seems to me that after linking to the ACS data, you would have had the option of constructing standard errors using the ACS replicate weights. Is that something that you considered?

Methodology on Creating the U.S. Linked Retail Health Clinic (LiRHC) Database

- Goal: Describe the creation of a new linked retail health clinic data base, and report on geographic distribution
- Methods & Data: Linked data (2018-2020) from three sources:
  - Convenient Care Association Membership - RHCs and other walk-in health care centers.
  - County Business Patterns Business Register (Census) - information related to payroll, employment size, and business characteristics.
  - National Plan and Provider Enumeration System (CMS) - data on individual health care providers and non-individual organizations.

Methodology on Creating the U.S. Linked Retail Health Clinic (LiRHC) Database

- Linking:
  - Employed a multistage matching process, whereby scores were given on different match criteria.
- Finding: Geographic Distribution
  - Almost half (47.0%) located in the South region, with majority (550 RHCs) found in the South Atlantic division.
  - Counts are consistent with other data sources.

Paper 2: Things I wondered about

- The paper notes that “Some CBPBR locations matching to the CCA records had an unexpected NAICS code from industries unassociated with retail trade locations or healthcare services.” Were there any lessons from these cases?
- Are there challenges to expanding the time component – would the matching have to be year by year?
- Would it be possible to obtain summary statistics on how well the matching process “worked”? For example, X% of CCA records matched to at least one site.
- Would it be possible to provide information about what questions this new linked data would be able to provide? For example:
  - Do the proportion of retail health clinics affect the price of medical service or the proportion of expenditure?
  - What is the likelihood of RHC co-located with other health facilities? Are they adding to or replacing other providers?

The Demographics of the Recipients of the First Economic Impact Payment

- Goal: To report on the demographics of receipt of the First Round Economic Impact Payments authorized under the CARES Act.
- Methods: Report on differences in the time of receipt by age, gender, income, number of children, income level

The Demographics of the Recipients of the First Economic Impact Payment

- Data
  - IRS records on receipt linked to Census data
- Findings
  - Consistent with IRS operational decisions, lower income individuals and families with children received payments earlier than higher income or families without children,
  - While there was a near 90 percent receipt rate for most racial/ethnic subgroups – it was highest for White individuals and lowest for Hispanic and Some Other Race individuals.

Paper 3: Demonstration of the Value of Linked Data

Lack of data not essential to administrating the program

- Linking with the census data allowed to see the demographics of the receipt of the credit.

If the data is on program participants – what is the counterfactual

- I'm not sure that the "control group" language holds, but the study used the demographics within the program to develop comparison groups – without this it would not have been possible to understand equity issues in the timing of credits.

Paper 3: Things I wondered about

- Not surprisingly, non-filers were the most likely to not receive a return within the first week by a wide margin (only 2 percent). This made me wonder of whether any other demographic gaps measured in the data could be explained by the percentage of non-filers in that group.
- In the paper, the authors state that “Many non-filers had no filing obligation in those years.” because of low income. Another reason for non-filing could be tax avoidance for higher income individuals.
- The authors present information on the percentage of eligible individuals that received the payment – it would also be interesting to see whether there were individuals – based on links to survey data – that would be estimated to be ineligible and received the payment.

Conclusion

- All papers provided evidence of the value of linking administrative data with survey data –performing analyses that would not have been possible without the linked data.
- The linked data also provided important information on the demographics of program participants.
- I'd encourage researchers to consider how "generalizable" results are to other similar programs

Methodology on Creating the U.S. Linked Retail Health Clinic (LiRHC) Database

- Retail health clinics (RHCs) are a relatively new type of health care setting and understanding the role they play as a source of ambulatory care in the United States is important. To better understand these settings, a joint project by the Census Bureau and National Center for Health Statistics used data science techniques to link together data on RHCs from Convenient Care Association, County Business Patterns Business Register, and National Plan and Provider Enumeration System to create the Linked RHC (LiRHC, pronounced “lyric”) database of locations throughout the United States during the years 2018 to 2020. The matching methodology used to perform this linkage is described, as well as the benchmarking, match statistics, and manual review and quality checks used to assess the resulting matched data. The large majority (81%) of matches received quality scores at or above 75/100, and most matches were linked in the first two (of eight) matching passes, indicating high confidence in the final linked dataset. The LiRHC database contained 2,000 RHCs and found that 97% of these clinics were in metropolitan statistical areas and 950 were in the South region of the United States. Through this collaborative effort, the Census Bureau and National Center for Health Statistics strive to understand how RHCs can potentially impact population health as well as the access and provision of health care services across the nati

The Demographics of the Recipients of the First Economic Impact Payment

Starting in April 2020, the federal government began to distribute Economic Impact Payments (EIPs) in response to the health and economic crisis caused by COVID-19. More than 160 million payments were disbursed. We produce statistics concerning the receipt of EIPs by individuals and households across key demographic subgroups. We find that payments went out particularly quickly to households with children and lower-income households, and the rate of receipt was quite high for individuals over age 60, likely due to a coordinated effort to issue payments automatically to Social Security recipients. We disaggregate statistics by race/ethnicity to document whether racial disparities arose in EIP disbursement. Receipt rates were high overall, with limited differences across racial/ethnic subgroups. We provide a set of detailed counts in tables for use by the public