Conducting ECE research using administrative data

(i.e. “big enough” data)

Wladimir Zanoni

Chapin Hall at The University of Chicago
Fellow. Harris School of Public Policy

December, 2015

Contact email:
wzanoni@chapinhall.org

This Talk:

Administrative databases are comprehensive:

- Measures they collect,
- Longitudinal with respect to the time frame
- Follow broad populations.

Those attributes make them attractive for ECE research

- Use of “big enough data” for ECE research has grown.

This Talk:

Nonetheless, extent of ECE research with administrative data is limited.

- High “barriers to entry”.
- Hard to assess advantages, disadvantages and potential complementarities with other data sources.

Aim:

Stimulate discussion about the opportunities and challenges involved in the use of administrative data.

This Talk:

1. **Basic elements of administrative data research:**
    - The supply of data and its quality
    - Research designs
    - Data linkage

2. **Research example:**
    - Childcare subsidies, parental earnings and child development

Basics of administrative data research:

- **The supply of data and its quality:**
  - Clerical work and supervision have incentives to collect data
  - Accuracy of data fields relates to the objectives of the agency collecting the data:
    - Missing values, errors and omissions:
      - Vary across agencies
      - Vary within agency across fields

Political economy of administrative data!

Basics of administrative data research:

- **Research designs:**

Conditional on having a research question defined researchers would have to ask:

1. Sample or population frame?
2. Cross sectional or longitudinal?
    - Prospective vs. retrospective
3. Single vs. multiple databases

Shared attributes with survey research methods!

Basics of administrative data research:

- **Administrative data linkage**
  - Individual level
    - De-duplication
    - Across- databases linkage
    - Methods for data linkage
    - Software
  - Common unit data linkage
  - Geography-based data linkage

Probabilistic Record Linkage

- Are pairs of data records associated to the same entity?
  - The “entity” is and individual and/or the family, school, ECE provider, neighborhood
  - Data records might or not be restricted to a single field (SSN; name; last name; DOB, etc.). There is an optimal number of fields
  - *Within the same database:* deduplication
  - *Across databases:* linkage

Individual level: Probabilistic Record Linkage

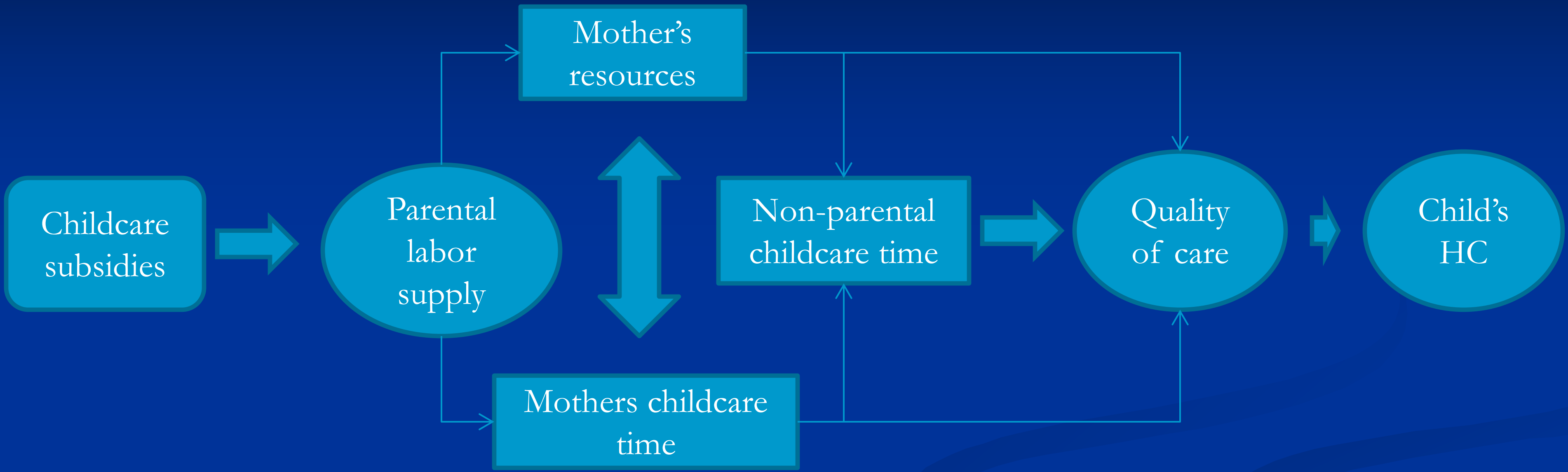
- Probabilistic record linkage:
  - A family of record linkage techniques
  - Constructs an “score” to pairs of records
  - Assigns unique identifiers to all pairs of records that score above a certain threshold.
  - Often neglected: moves the real from population-base to probabilistic
    - With unexplored implications for precision and/or bias

This Talk:

2. Research example

- Childcare subsidies, earnings, and child development

Research question: how do childcare subsidies affect child development?



1) Childcare subsidies affects labor supply

2) Mother's own resources and time invested in HC are reallocated

3) Parental care is substituted with non-parental care of certain quality

4) Quality of care affects development of child's human capital

The policy problem:

How to balance intergenerational effects of childcare subsidies on economic success and income inequality?

Today: Employment probability, and earnings

Tomorrow: Skills that characterize young labor force

Challenges to answer the research question:

1. **We have to rely on observational data**

2. **Limited availability/quality of data:**
   - For this specific research purpose ECLS-K/ Three-city

3. **Pervasive endogeneity-selection problems:**
   - Unobservable factors explain simultaneously take up and cognitive development

   - Women do not select into childcare subsidies at random

How do we address the challenges?:

We built a unique dataset to identify effects:

- Chapin Hall’s administrative data resources
    - CPS 3rd graders in school years 2006-2010
    - Matched to Food Stamps recipients
    - Matched to childcare subsidy recipients (CCDF)
    - Matched to Unemployment Insurance (quarterly wages)
    - Addresses “geocoded” and 2000 census data imputed

A unique dataset:

CPS 3rd grade Students

SID

CPS:
- Dates of birth
- Names
- Language
- Ethnicity
- Disability
- Cared by mother
- ISAT/ITBS
- School characteristics
- Address

2000 CENSUS:
- Demographics
- Community area
- Census tract
- Census block group

CHDHSID

CHDHSID

CHDHSID

CHDHSID

CHDHSID

FOOD STAMPS CASE:
- Dates of birth
- Number of people
- Education (TANF)
- Gender

MOTHER

Age range (16-40)

CHMSID

UNEMPLOYMENT INSURANCE:
- Quarterly earnings

INDIVIDUAL_ID

CHILDCARE TRACKING SYSTEM:
- Take up
- Amount
- Type
- Provider's address
- User's addresses

Geographical Density:
- Users
- Providers

A unique dataset:

- ISAT/ITBS TEST SCORES (1991-2010)

- HISTORY OF CHILDCARE PARTICIPATION (1997-2010)
  - MONTHLY TAKE UP
  - TYPE OF PROVIDER
  - AMOUNT RECEIVED

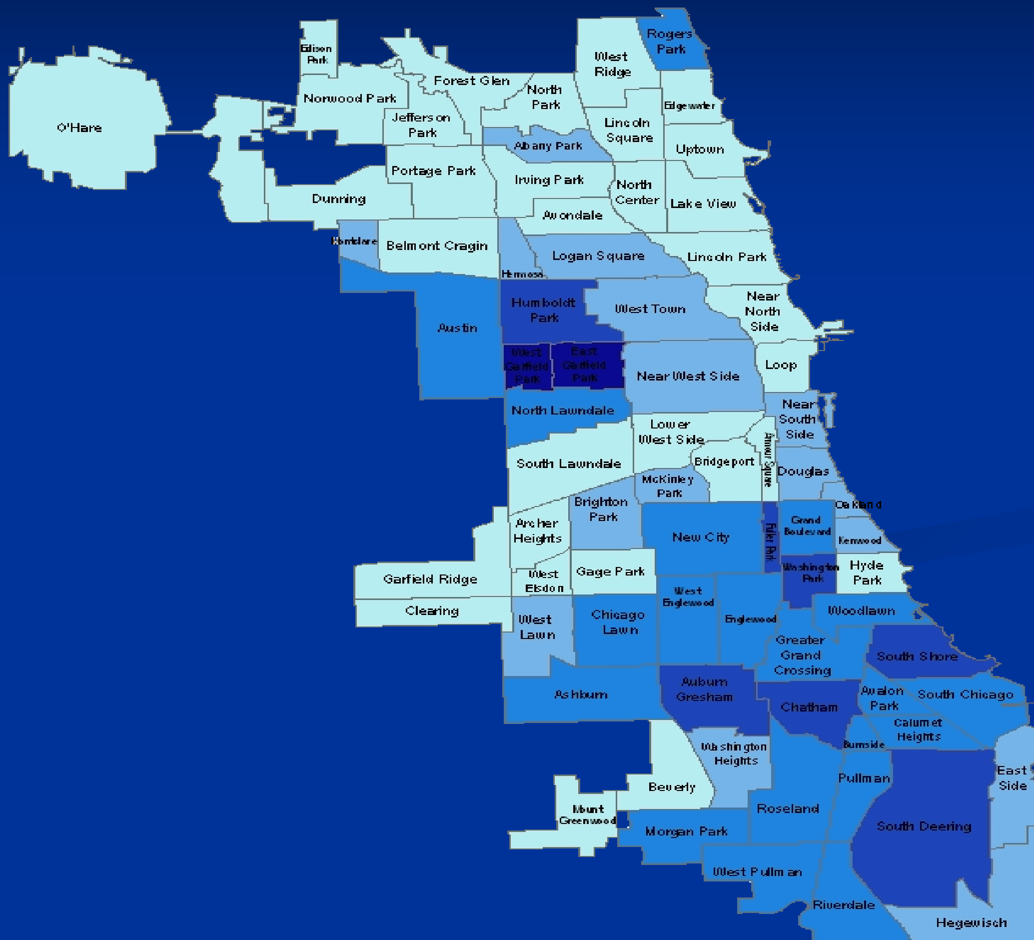
- 11 YEARS OF QUARTERLY EARNINGS (1995-2006)

- DEMOGRAPHIC CHARACTERISTICS

- TRACK OF GEOGRAPHIC LOCATION (1990-2000 CENSUS)
  - ENABLES CENSUS DATA MATCH

The Data: Chicago: Density in utilization

77 Community areas (March 2009)



Deeper blue implies higher density in use of childcare subsidies

Does take up increases monotonically with density?

Empirical Analysis:

OLS AND IV-2SLS ESTIMATES OF THE SUBSIDY EFFECT
(SUBSIDIES TAKEN ANYTIME IN THE 0-5 YEARS OLD AGE RANGE)

	(1)	(2)	(3)	(4)
MATH				
OLS coefficient	**-0.0042****	**-0.0041****	**-0.0042****	**-0.0042****
Standard error	(0.0018)	(0.0018)	(0.0018)	(0.0018)
IV- treatment coefficient	**-0.1083***	**-0.0394**	**-0.0769**	**-0.0149**
Standard error	(0.0618)	(0.0743)	(0.0679)	(0.0803)
F-statistics (1st stage)[a]	25.32	11.82	20.04	10.34
READING				
OLS coefficient	**-0.0044****	**-0.0041****	**-0.0043****	**-0.0043****
Standard error	(0.0019)	(0.0019)	(0.0019)	(0.0019)
IV- treatment coefficient	**-0.1600****	**-0.0547**	**-0.1115**	**-0.0063**
Standard error	(0.0681)	(0.0782)	(0.0727)	(0.0841)
F-statistics (1st stage)[a]	24.66	11.45	19.61	10.13
ITBS-VARIABLE[b]	NO	YES	NO	NO
PRED. VARIABLES[c]	NO	YES	YES	YES
IV DENSITY[d]	YES	YES	YES	YES
IV Comm. Area ITBS[e]	NO	NO	NO	YES

Column (4): instrument CT ITBS with CA ITBS

Conclusion

- There are advantages
- There are disadvantages
- There are complementarities
- Advocacy efforts are needed.

Thanks!

Contact email:
wzanoni@chapinhall.org
wzanoni@uchicago.edu