Developing job linkages for the Health and Retirement Study

John Abowd, Margaret Levenstein, Kristin McCue, Dhiren Patki, Ann Rodgers,

Matthew Shapiro, Nada Wasi

NCRN Spring 2016 Meeting, May 10, 2016

Acknowledgements and disclaimers

This research is supported by the Alfred P. Sloan Foundation through the Census-HRS project at the University of Michigan with additional support from the Michigan Node of the NSF-Census Research Network (NCRN) under NSF SES 1131500.

This research uses data from the Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by the following National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

Outline

- Background on HRS, CenHRS
- Approach to linkage
- Work using a small set of HRS jobs
- Some preliminary results
- Challenges

HEALTH AND RETIREMENT STUDY

A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

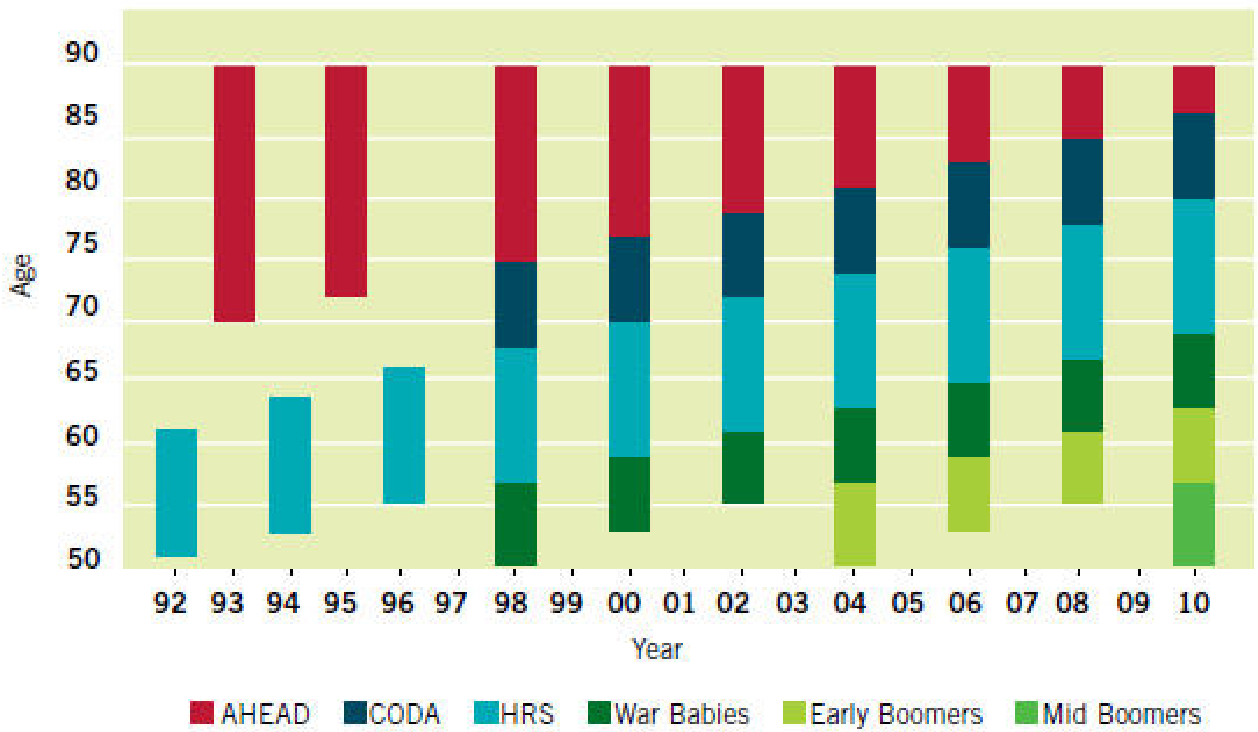
37,000 + Americans over the age of 50

- Surveyed every two years since 1992
- New cohorts added in 1993, 1998, 2004, 2010, 2016
- Includes both spouses
- Oversamples minorities
- Follows respondents through death

HEALTH AND RETIREMENT STUDY

A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

THE HRS LONGITUDINAL SAMPLE DESIGN



Census-Enhanced HRS

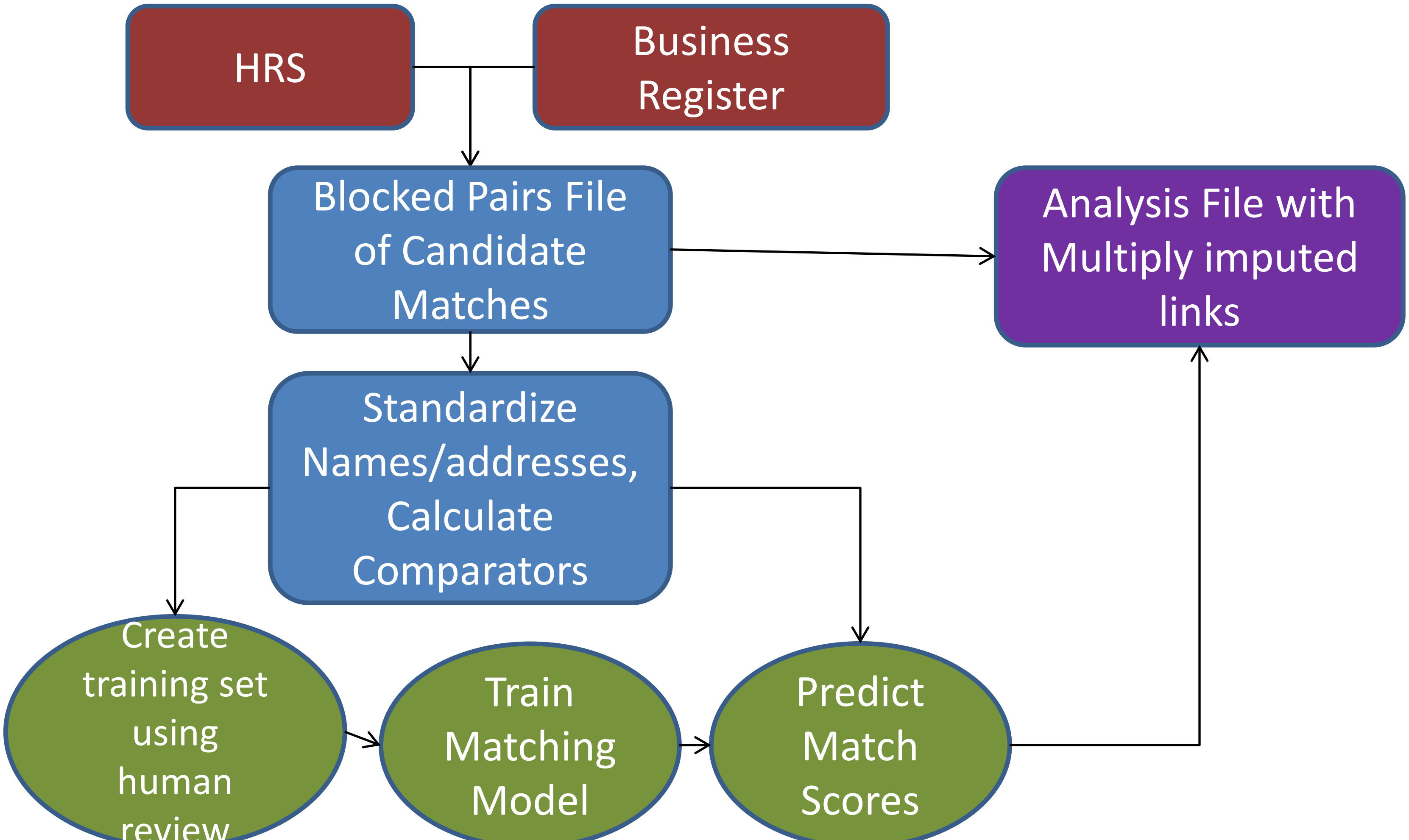
UMichigan/Cornell/Census collaboration

Goal: New info on HRS respondents in employer and co-worker context

Develop new data infrastructure:

- HRS-BR Crosswalk
- New measures of employer characteristics
- Enhance HRS public-use datasets

Linkage Process Flow



First steps:

- Use a subset of 1992 HRS private-sector jobs, 1992 BR to work out methods
- Block on:
  - 10-digit phone number, where possible
  - 3-digit zip code, otherwise
- Standardize address and name fields, using rules developed specifically for business names
- Compute Jaro-Winkler string comparator scores for names and addresses

Construct set of pairs

- 1,232 1992 HRS jobs from 7 states
- Exclude if missing employer name or state, or missing both zip3 and phone # (10%)

- <10% of phone numbers successfully blocked
- Almost always at least 1 BR entry in zip3 block

Initial set of blocked pairs

- All possible within-block pairs = 18.3M
- JW scores comparing name, address
- Stratify using 4x4 cross-classification of JW scores
- Mean pairs per sampled HRS job=3,100, but varies from 1 to 20,000 across bins.
- Lowest JW scored bin accounts for:
  - 98% of pairs blocked on 3-digit zip
  - 42% of those blocked on 10-digit phone number

Creating training set

- Sample 100 pairs from each stratum
- Each sampled pair reviewed by >=2 reviewers
- Reviewers see 1 pair at a time
- Assign separate scores for firm, establishment
- Score as follows:

1	=	Yes, match
2	=	Probably match
3	=	Maybe-maybe not
4	=	Probably not match
5	=	Not match
6	=	Not enough information

Results of review

- 3,400 reviews, 7 reviewers

Match?	Establishment	Firm
Yes	10%	18%
Maybe	13%	11%
No	76%	71%
Not enough info	<1%	<1%

- Disagreement across reviewers:
  - 5% for yes/no reviews
  - 63% for maybe/not enough info
- Use only yes/no reviews in estimating model (3,100)

Match rates by blocking factor

Share of reviews scored as match		
Blocked on	Establishment	Firm
10-digit phone number	94%	100%
3-digit zipcode	11%	19%
Note: Reviews scored Probably match, maybe/maybe not, probably not match, or not enough information are excluded from denominator.		

Modeling approach

Model propensity for record from HRS to match record from the BR

- Estimate model parameters using training set
- Calculate agreement probability for all possible pairs within block

Multiply impute links using agreement probabilities

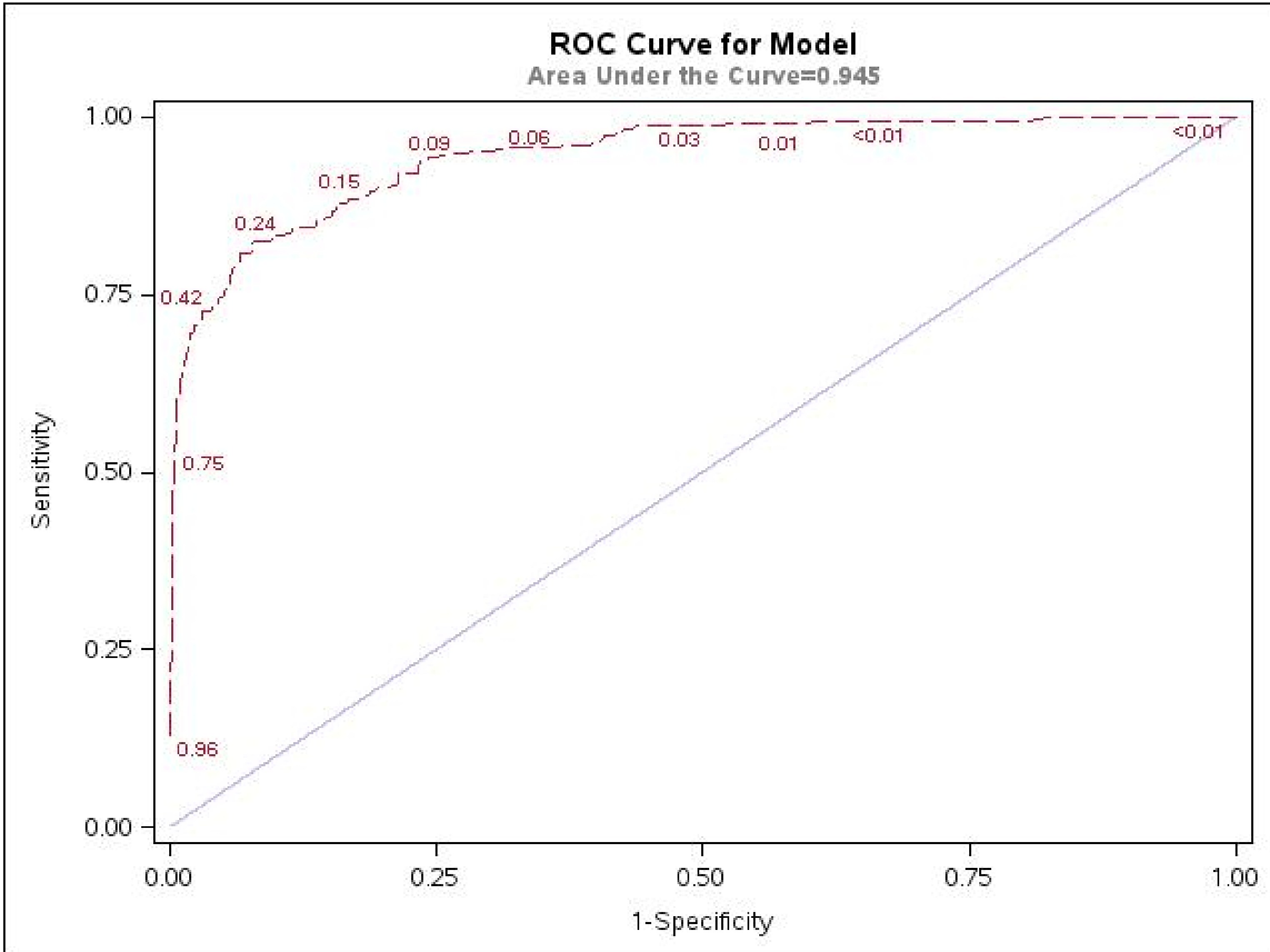
Training our matching model

- Using logistic model: dep var = 1 if pair is scored as a match, 0 otherwise
- Regressors: splines of continuous variables, indicators, and a full set of interactions
- To limit overfitting and to minimize out of sample error, we use elastic net shrinkage (Zou and Hastie, 2005)
  - Elastic net shrinkage reduces the dimensionality of the covariate vector
  - Idea: the optimal set of covariates is chosen to minimize cross-validated test error

Available model covariates

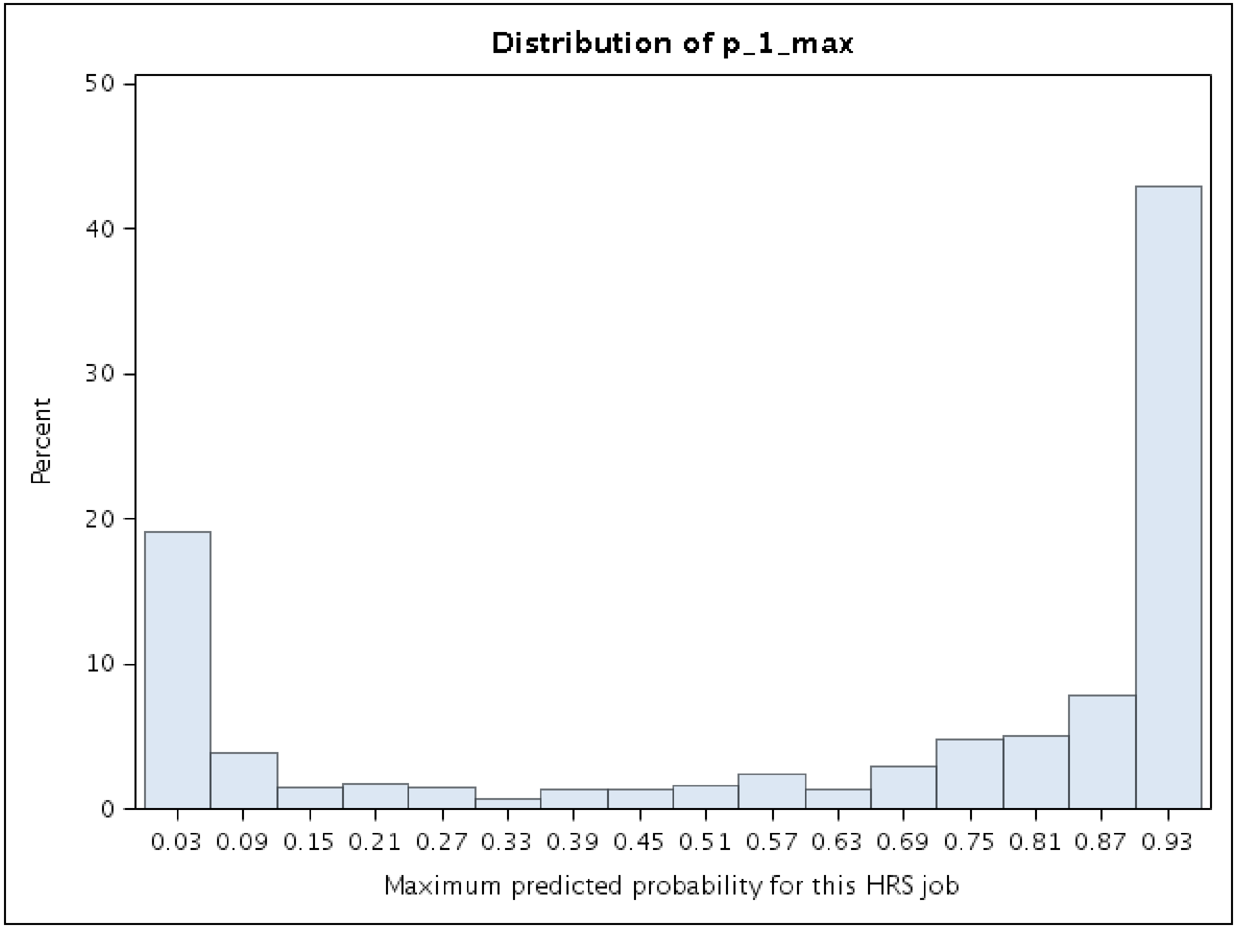
- JW scores for agreement of name, address fields
- Employment for establishment/employer for categories: 0/missing, 1-4, 5-14, 15-24,25-99, 100-499 500+
- Agreement on 3-digit, 5-digit zip code
- Agreement on industry—2 digit SIC
- Whether BR record is for single- or multi-unit
- Whether HRS employer offers health insurance/pension
- Business density—number of establishments in tract or per square mile

True positive rate



False positive rate

Distribution of maximum predicted probability using only JW scores



Challenges

- What to do when block does not include any high probability matches?
- Possible reasons
  - Blocking strategy excluded correct match
  - Blocking didn't fail:
    - Model failure
    - HRS information too garbled to support matching