

Data Linkage in the Survey of Occupational Injuries and Illnesses

Ellen Galantucci

Research Mathematical Statistician
Office of Compensation and Working Conditions
Bureau of Labor Statistics

November 3, 2021



Outline

- ▶ Description of the Survey of Occupational Injuries and Illnesses (SOII)
- ▶ SOII linkage with Occupational Safety and Health Administration (OSHA)
 - ▶ Probabilistic
 - ▶ OSHA Injury Tracking Application (ITA) Identification Number
 - ▶ Employer Identification Number
- ▶ Uses of linked data
 - ▶ Burden reduction
 - ▶ Imputation
 - ▶ Non-response bias analysis

Survey of Occupational Injuries and Illnesses

- ▶ The Survey of Occupational Injuries and Illnesses (SOII) program has collected injury and illness information from business establishments each year since 1972
- ▶ Currently, the survey uses an annual sample of approximately 230,000 establishments and has a response rate of about 89%
- ▶ The survey covers wage and salary earners and some volunteers across all 50 states, D.C., and three U.S. territories in most industries
- ▶ While some types of establishments are excluded, such as agricultural establishments with fewer than 10 employees (NAICS 111 and 112) and the self-employed, the survey covers the vast majority of the United States economy



Survey of Occupational Injuries and Illnesses

- ▶ Collects summary information, such as total number of cases resulting in days away from work or days of job transfer and work restrictions
- ▶ Also collects detailed case and demographic information about some injury or illness cases



Occupational Safety and Health Administration (OSHA)

- ▶ In 2016, OSHA issued a rule that would require certain establishments to submit summary case information on an annual basis
- ▶ Data would be collected through their Injury Tracking Application (ITA)
- ▶ Applies to all business establishments with 250 or more employees in industries required to maintain OSHA logs and all establishments with 20-249 employees in industries that have historically high rates of occupational injuries and illnesses

Office of Management and Budget (OMB)

- ▶ Two agencies within the Department of Labor were now collecting the same information
- ▶ This was an increased burden on establishments that had to submit the same information to both agencies
- ▶ The overlap in reporting requirements was about 40% of the SOII sample
- ▶ OMB tasked us with finding a way to reduce burden



Probabilistic Linkage

- ▶ The two datasets lack good quality, unique identifiers
- ▶ Variables such as establishment name, address, and employment totals are noisy
- ▶ The BLS hired contractors to attempt to link establishments in the two datasets



Probabilistic Linkage

- ▶ The contractors created a probabilistic linkage methodology based on Fellegi-Sunter (1969)
- ▶ Establishment name, address, industry, employment total, phone number, and email addresses domain used for matching
- ▶ Had to standardize variables using methods such as geocoding
- ▶ Linked the OSHA ITA data to the Quarterly Census of Employment and Wages (QCEW) rather than the SOII data



Some Concerns with Probabilistic Linkage Methodology

- ▶ Level of blocking (zip code) reduced size of potential matches but eliminated some accurate pairs
- ▶ Challenges accounting for addresses that housed multiple establishments
- ▶ Geocoding difficulties
- ▶ Weighting based on how often matches occurred within the datasets, which meant variables like employment not seen as important
- ▶ QCEW and OSHA both have multiple fields for establishment names



Addition of Collection of OSHA ITA ID and EIN

- ▶ SOII began collecting OSHA ITA IDs at the end of 2018 data collection
- ▶ OSHA began collecting EINs during 2019 data collection
- ▶ These two variables would provide better identifiers that could be matched but would need to be validated



Collection of OSHA ITA ID

- ▶ 6,388 establishments provided a numerical, non-zero OSHA ITA ID during 2018 data collection
 - ▶ The number of establishments that provide this increased substantially in 2019 data collection
 - ▶ 221 establishments submitted their industry classification (NAICS code)
 - ▶ 292 submitted their Internet Data Collection Facility (IDCF) identification number which is used for logging in to the SOII data collection instrument
 - ▶ 399 additional establishments submitted a number that had no match in the OSHA database
- ▶ 5,476 remaining submitted numbers were potential matches



Variables Considered for Validating OSHA ITA ID

- ▶ Employment totals within 30%
 - ▶ Required for match
 - ▶ About 90% met this criterion
- ▶ NAICS code and zip code
 - ▶ About 50% of those that matched on employment matched on NAICS code and zip code
 - ▶ About 25% of those that matched on employment matched on zip code but not NAICS code
 - ▶ About 15% of those that matched on employment matched on NAICS but not zip code

Validating Accuracy of Matches

	NAICS/zip matched	Zip only matched	NAICS only matched	Neither NAICS nor zip matched
DAFW and DITR both matched	2,255 (89.8%)	1,088 (90.7%)	759 (90.9%)	316 (81.0%)
Total case counts matched but did not match on both DAPW and DITR	115 (4.6%)	56 (4.7%)	28 (3.4%)	18 (4.6%)
DAFW matched but not DITR or total case count	50 (2.0%)	27 (2.3%)	19 (2.3%)	12 (3.1%)
DITR matched but not DAPW or total case count	56 (2.2%)	17 (1.4%)	15 (1.8%)	20 (5.1%)
None of the above matched	35 (1.4%)	11 (0.9%)	12 (1.4%)	24 (6.2%)
Total establishments	2,511	1,199	835	390

Validating Accuracy of Matches

- ▶ If Days Away from Work and Days of Job Transfer and Work Restriction counts both matched, it was considered a match
- ▶ If the total case count matched, it was considered a match
- ▶ Everything else was manually validated using a combination of establishment name, address, and name of the person who submitted the data
 - ▶ For every unit that matched on NAICS and/or zip code, the company matched though the specific establishment could not always be confirmed as correct
 - ▶ For remaining units, 17/56 appear to be the same company
- ▶ The combination of employment and either NAICS code or zip code appears to be sufficient to validate the accuracy of the OSHA ITA ID matches



Matching by Employer Identification Numbers (EIN)

- ▶ EINs are not unique to establishments - they are unique to companies
- ▶ Despite that, using other variables such as address, employment totals, and establishment names can help to match establishments between SOII and OSHA
- ▶ In some cases, there are multiple establishments with the same EIN in close proximity, which are nearly impossible to differentiate between



Uses for linked data

- ▶ Reduce burden in SOII
 - ▶ BLS now has an API to collect data from OSHA
 - ▶ Any units that have already reported to OSHA and can be matched do not have to report summary information to SOII
- ▶ Imputation or sample supplementation in SOII
 - ▶ We are brainstorming ways to use the data from OSHA to fill in gaps that exist within the SOII data
 - ▶ The challenge is that the data have not had the same data quality screening as normal SOII data has



Uses for linked data

- ▶ Non-response bias analysis
 - ▶ Using EIN in combination with other variables, I impute the data of SOII non-respondents
 - ▶ I use those imputations to predict injury and illness counts by industry
 - ▶ This analysis has shown that SOII data do not suffer from non-response bias



CONTACT INFORMATION

Ellen Galantucci, Ph.D.
Research Mathematical Statistician
Statistical Methods Group
Office of Compensation and Working Conditions
Bureau of Labor Statistics
galantucci.ellen@bls.gov

