



September 2, 2022

Introducing the Medical Expenditure Panel Survey-Insurance Component with Administrative Records (MEPS-ICAR): Description, Data Construction Methodology, and Quality Assessment

Thomas A. Hegland, PhD, Alice Zawacki, PhD, and G. Edward Miller, PhD*

**Hegland [corresponding author] is an economist at the Agency for Healthcare Research and Quality; thomas.hegland@ahrq.hhs.gov. Zawacki is a senior economist at the United States Census Bureau; alice.m.zawacki@census.gov. Miller is a senior economist at the Agency for Healthcare Research and Quality; ed.miller@ahrq.hhs.gov. We would like to thank Kristin McCue, Danielle Sandler, and John Voorheis from the U.S. Census Bureau's Center for Economic Studies for sharing their expertise with Census Bureau and administrative records data.*

Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily reflect those of the Agency for Healthcare Research and Quality, the Department of Health and Human Services, or the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release. Disclosure Review Board Approval Numbers CBDRB-FY22-047 and CBDRB-FY22-292; DMS project number 7514872.

Abstract

This report introduces a new dataset, the Medical Expenditure Panel Survey-Insurance Component with Administrative Records (MEPS-ICAR), consisting of MEPS-IC survey data on establishments and their health insurance benefits packages linked to Decennial Census data and administrative tax records on MEPS-IC establishments' workforces. These data include new measures of the characteristics of MEPS-IC establishments' parent firms, employee turnover, the full distribution of MEPS-IC workers' personal and family incomes, the geographic locations where those workers live, and improved workforce demographic detail. This report details the methods used for producing the MEPS-ICAR. Broadly, the linking process begins by matching establishments' parent firms to their workforces using identifiers appearing in tax records. The linking process concludes by matching establishments to their own workforces by identifying the subset of their parent firm's workforce that best matches the expected size, total payroll, and residential geographic distribution of the establishment's workforce. The report presents statistics characterizing the match rate and the MEPS-ICAR data themselves. Key results include the fact that match rates are consistently high (exceeding 90 percent) across nearly all data subgroups, and that the matched data exhibit a reasonable distribution of employment, payroll, and worker commute distances relative to expectations and external benchmarks. Notably, employment measures derived from tax records, but not used in the match itself, correspond with high fidelity to the employment levels that establishments report in the MEPS-IC. The construction of the MEPS-ICAR dataset significantly expands the capabilities of the MEPS-IC, and presents many opportunities for analysts.

Suggested Citation

Hegland, T., Zawacki, A., and Miller, E. *Introducing the Medical Expenditure Panel Survey-Insurance Component with Administrative Records (MEPS-ICAR): Description, Data Construction Methodology, and Quality Assessment*. Methodology Report #35. September 2022. Agency for Healthcare Research and Quality, Rockville, MD.
http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr35/mr35.pdf

* * *

The estimates in this report are based on the most recent data available at the time the report was written. However, selected elements of Medical Expenditure Panel Survey (MEPS) data may be revised on the basis of additional analyses, which could result in slightly different estimates from those shown here. Please check the MEPS website for the most current file releases.

Table of Contents

I. Introduction	1
II. Description of the MEPS-ICAR	4
Firm/Worker-Level Data	4
Establishment/Worker-Level Data	5
Firm- and Establishment-Level Data	6
A Choice of Employment Concepts: Over the Year vs. Point in Time	7
Limitations	8
III. MEPS-ICAR Construction Methodology	9
Stage 1: Prepare the MEPS-IC and Administrative Record Data	9
Stage 2: Produce the Firm-Level Link and Prepare for Forming Establishment-Level Links	10
Stage 3: Match Workers in MEPS-IC Firms to MEPS-IC Establishments	12
Case 1: Single-Establishment Firm.....	12
Case 2: One Establishment from a Multi-Establishment Firm	12
Case 3: Multiple Establishments from a Multi-Establishment Firm.....	14
Stage 4: Finalize the Match and Identify Failed Matches.....	15
Stage 5: Produce Turnover Statistics and Point-in-Time Weights	16
IV. Assessing the MEPS-ICAR	18
Establishment- and Firm-Level Statistics	18
Match Rates.....	18
Employment and Turnover	19
Payroll	21
Employment and Payroll by Single-Establishment vs. Multi-Establishment Firm Status.....	22
Alternative Metrics on Match Quality	23
Worker-Level Statistics.....	24
Worker-Level Matches and Demographic Characteristics.....	24
Full Distributions of Worker Ages, Wages, and Family Incomes.....	26
Commuting Distances for Workers	27
V. Conclusion.....	29
VI. References	31
VII. Tables	33

Introducing the Medical Expenditure Panel Survey-Insurance Component with Administrative Records (MEPS-ICAR): Description, Data Construction Methodology, and Quality Assessment

Thomas A. Hegland, PhD, Alice Zawacki, PhD, and G. Edward Miller, PhD

I. Introduction

This methodological report details the construction of a new dataset that considerably expands the analytical scope of the Agency for Healthcare Research and Quality's Medical Expenditure Panel Survey-Insurance Component (MEPS-IC) by linking it to Decennial Census data and to Internal Revenue Service (IRS) administrative tax records drawn from W2 forms and Form 1040s. Our newly constructed dataset, the MEPS-IC with Administrative Records (MEPS-ICAR), is the first and only nationally representative dataset of United States businesses that both characterizes businesses' health benefits packages and offers detailed socioeconomic information about these businesses' workers and their families. As such, this new dataset should enable analysts to improve our understanding of a range of issues, including how employers' health benefits packages vary with their workforce's personal and family characteristics, how employers make decisions relating to the tradeoff between offering more generous health benefits and higher wages, and how various state and federal policies that target individuals and families (e.g., Medicaid expansions) affect these peoples' employers and their health benefits–related decisions.

The MEPS-ICAR currently spans 2005–2017, excluding 2007, and a planned update will extend the data through 2020 in the near future, with further annual updates being planned as well.¹ The MEPS-ICAR's business establishment data is primarily derived from the MEPS-IC, which collects detailed information from a sample of private sector establishments about their health insurance benefits packages, along with some additional summary statistics characterizing the establishment, its parent firm (when the establishment is part of a multi-establishment firm), and, to a less detailed extent, the establishment's employees (AHRQ, 2005–2017; Davis, 2018).^{2,3} The MEPS-ICAR adds to this data the ability to observe the full distribution of wages paid by each MEPS-IC establishment to everyone it employed in each year, as well as observation of each linked worker's family income, family size, age,

¹ The year 2007 is excluded because the MEPS-IC was not conducted for reference year 2007, due to the transition from retrospective to current-year data collection.

² The MEPS-IC also collects annual data on employee health insurance benefits offered by approximately 3,000 state and local sampled governments, but these units are not included in these linkages.

³ Single-establishment firms are generally referred to as "single-units" (SUs) and multi-establishment firms as "multi-units" (MUs) in Census Bureau documentation.

race, ethnicity, sex, marital status, family composition, and geographic location of residence. The MEPS-ICAR also offers all of this information for each MEPS-IC establishment's parent firm and its workforce, alongside some additional information on each establishment's (and its parent firm's) annual employee turnover rate, annual total number of hired workers, and annual total number of separated workers. Further detail on these new measures available in the MEPS-ICAR is given in Section II of this report.

Glossary

Establishment. A particular physical location where business activity takes place. Ex: the ice cream shop located at the corner of 5th and main; a company's corporate headquarters.

Firm. A business as a whole. A firm may own or operate multiple establishments. The firm owning or operating a given establishment is known as its parent firm.

Building the MEPS-ICAR proved to be a complex task. While IRS tax records do indicate which firm employs each worker on W2 forms, they do not record the particular establishment at which each worker is employed. Even though pre-existing identifiers available from the Census Bureau and the IRS can be used to link MEPS-IC establishments to their parent firms and those parent firms to their workforces, there is no direct way to link particular MEPS-IC establishments to just their own employees, apart from single-establishment firms. For any given MEPS-IC establishment that is part of a multi-establishment firm, we link it to its workforce by searching among all employees of its parent firm and assigning to it a collection of workers that (a) contains a number of workers as close as possible to the number we expect to find for the establishment,⁴ (b) reports a total amount of W-2 wages that is as close as possible to the establishment's expected payroll total, and (c) has an average commute distance between the establishment's location and each worker's home residence that is as low as possible. Finding a collection of workers satisfying these conditions is a difficult combinatorial optimization problem for which it is computationally infeasible to provide an exact solution.⁵ Nevertheless, our chosen approach for approximating a solution to this has advantageous properties and allocates workers to establishments in a manner such that any assignment errors are mitigated by the similarity of any potential alternative assignments: if a worker is erroneously linked to an establishment when a different worker should have been linked instead, the two workers will still be employed by the same parent firm, typically should be similarly paid, and typically should have residences near each other. We also impose match quality standards that reject any poor-quality establishment-worker linkages that may result from this process. An overview of our methodology for constructing the MEPS-ICAR,

⁴ Notably, this number is not, in general, the number of employees an establishment reports having in the MEPS-IC survey. This is because establishments report steady-state employment levels to the MEPS-IC, while in the IRS data, they should be matched to every individual employed by the establishment over the course of the year. These two numbers are only the same if the establishment does not have any employee turnover.

⁵ In addition to difficulties arising from possible measurement error in the input data, the computational infeasibility of finding exact solutions to this problem derives from its close relationship to NP-hard problems like the knapsack problem and the generalized assignment problem.

including more details on the algorithms we use for making the establishment-worker match, is available in Section III of this report.

In terms of match performance, we succeed in linking 92.89 percent of MEPS-IC establishments to their workforces. These linked establishments capture about 93.33 percent of employment in the MEPS-IC, as measured by the employment levels establishments self-report on the survey. This match rate is consistently high across most data years, geographic areas (i.e., Census divisions), industries, firm size categories, and establishment size categories. Notably, successful match rates for establishments that are members of multi-establishment firms are broadly similar to those of establishments that are members of single-establishment firms, despite the greater difficulty associated with matching the former type of establishment. The largest exception to the tendency of match rates to exceed 90 percent across most data subgroups is the case of establishments that represent single-employee businesses, where our match rate dips to 87.80 percent. This lower match rate likely reflects a mixture of difficulty tracking micro-firms and possible differences in how these businesses file taxes relative to other businesses. Section IV of this report presents further detail on these match rates by subgroup.

In addition to examining match rates, we also consider a suite of statistics characterizing the matched workforces and comparing them to (1) information reported by MEPS-IC establishments about their workforces and (2) external data sources. To highlight a few key findings, first, we find that the number of workers matched to each establishment tends to hew quite closely to the number targeted by our matching procedure. The median establishment is matched to the number of workers targeted for it, while the mean establishment is matched to about 2.5 (or 8 percent) fewer workers than targeted. This close correspondence indicates that our establishment-workforce matching algorithm tends to successfully hit its targeted employment levels. Second, the "typical" employment levels that establishments are asked to report to the MEPS-IC survey tend to correspond quite closely (a difference of less than two employees, or 10 percent, at the mean establishment) with the steady-state employment levels implied by observing the number of an establishment's workers that, per tax records, remain employed at the establishment from one year to the next. Since this steady-state employment measure was not used in the match-making process, this correspondence represents a favorable external check on the quality of the match and its inputs (including the employment data that establishments report on the MEPS-IC survey itself). Third, the process also matches establishments to targeted payroll levels with a still high, but somewhat lower, level of fidelity, reflecting the fact that the match was generally written to prioritize employment levels. Fourth, match statistics such as the above generally are about equally favorable for establishments of single-establishment and multi-establishment firms, suggesting no reduction in match quality for establishments of this more difficult latter type. Finally, comparison of the distribution of commute distances observed between

workers and establishments in the MEPS-ICAR to the distribution of commute distances reported in the National Household Travel Survey (NHTS) indicates that the two distributions correspond quite closely, within the 90th and even 95th percentiles by commute distance. This result suggests that the MEPS-ICAR's minimization of commute distances broadly succeeded in producing a realistic commute distance distribution, thereby suggesting that this component of the algorithm also was helpful in forming correct worker-establishment assignments. The full set of match quality statistics are included, alongside the match rates, in Section IV of this report.

Finally, in Section V, we conclude by highlighting a selection of key areas of research likely to benefit from new data in the MEPS-ICAR. We also discuss certain benefits accrued to the baseline MEPS-IC project as a result of the MEPS-ICAR's construction.

II. Description of the MEPS-ICAR

Firm/Worker-Level Data

The MEPS-ICAR consists of a worker-level file containing, for each year, data on all individuals employed at firms with at least one establishment sampled in the MEPS-IC for that year, potentially observing workers more than once within the data when they work for multiple MEPS-IC employers throughout the year. The MEPS-IC survey data is collected from approximately 25,000–30,000 private sector establishments sampled from the 6.5–7.5 million contained in the Business Register (BR) frame that the U.S. Census Bureau maintains. The MEPS-IC contains a wide range of data from establishments about their health insurance benefits, including details on up to four offered plans and the number of workers electing to take up each plan. The linked administrative records are drawn from the full set of each year's IRS W2 Forms and Form 1040s, as well as from the 2000 and 2010 Decennial Censuses.

For every individual worker matched to a sampled MEPS-IC firm, we can observe their reported W2 data for their job at that firm, including wages and tips, Federal Insurance Contributions Act (FICA) wages, and the amount of deferred compensation.⁶ Using a set of pre-existing links provided by the Census Bureau, we link these workers to Decennial Census records to obtain their age, race/ethnicity, and sex (Wagner and Layne, 2014). We link 93.64 percent of MEPS-ICAR workers to at least one Decennial Census.

⁶ The W2 files available to Census Bureau researchers include the total amount for deferred compensation, but do not include data for the different types of deferred compensation, such as elective deferrals to a section 401(k) arrangement (Code D), a section 403(b) salary reduction agreement (Code E), a section 408(k)(6) salary reduction SEP (Code F), a section 457(b) deferred compensation plan (Code G), or a section 501(c)(18)(D) tax-exempt organization plan (Code H).

In addition to the data derived from the W2s and the Decennial Census, we link workers to Form 1040 information. We successfully match 91.61 percent of workers to a Form 1040, thereby offering us additional financial information for each worker's tax filing unit (which we will hereafter refer to as a worker's family).⁷ Specifically, we can observe family wage and salary income, taxable dividend income, taxable interest income, gross rent and royalty income, total money income, social security income, earned income, and tax-exempt interest income. Some limited information can also be derived about whether income is coming from a sole proprietorship, farming, an S-corporation, or self-employment. This Form 1040 information, in conjunction with W2 wages, allows us to calculate the share of a linked worker's family income derived from their job at a sampled MEPS-IC firm. Beyond income, Form 1040s also provide information about the worker's family structure,⁸ including the number of income-earning individuals, the number of dependents, the filer's marital status (as derived from the Form 1040's filing status), and exemptions that can be claimed for children and other dependents. Beyond these measures directly derived from Form 1040s, we also calculate the share of workers at the firm overall who can be linked to their Form 1040. One point of caution to bear in mind when considering Form 1040-derived data is that workers from lower-income families are not required to file a Form 1040, though some may file nevertheless in order to access certain refundable tax credits or for other reasons.⁹

Finally, we also have information about where workers live. Aggregating this information to the firm level allows us to observe the geographic extent of firms' workforces—information that can be important for understanding firms' exposure to various state-level policies (e.g., tax changes, Medicaid policies, health insurance regulations). We obtain this information about workers' residential locations from a mixture of sources: Form 1040s, the Decennial Censuses, and the Longitudinal Employer-Household Dynamics (LEHD) Residence Candidate File (Graham, Kutzbach, and Sandler, 2017).

Establishment/Worker-Level Data

While the firm/worker-level data described above consists of the broadest cut of the data in the MEPS-ICAR, we also link workers to specific establishments, or physical locations within each firm, that appear in the MEPS-IC sample. Overall, we link

⁷ A tax filing unit is the set of people that file together on an IRS Form 1040. This is not always the same as a worker's household or family, though we refer to tax filing units as families in the text because the family is the nearest general analogue to the tax filing unit concept.

⁸ A caveat for these data is that workers only report about their tax filing units on IRS Form 1040s—a concept that does not necessarily correspond precisely with their family unit reflected in other data sources.

⁹ As of 2021, the income threshold where filing becomes required for a single person under the age of 65 is \$12,550. For a married couple filing jointly where both spouses are under the age of 65, the threshold is \$25,100.

92.89 percent of MEPS-IC sampled establishments to workers.¹⁰ All variables available in the worker-firm linked portion of the MEPS-ICAR are also available at the worker/establishment level. Additionally, for workers linked to their establishment of employment, we have estimates of the distance from each worker's residence to the physical location where they report to work. These distances are derived from the worker's residential data discussed above, coupled with the establishment's physical location available in the MEPS-IC.

Firm- and Establishment-Level Data

The MEPS-ICAR worker-level dataset captures a very large number of employees and can be unwieldy simply due to its large size. For analytical convenience, these worker-level data have been rolled up to the firm and establishment level. These files offer sums, means, and percentiles of all worker-level variables calculated at the establishment and firm levels.

In addition to the above statistics, we also calculate establishment- and firm-level worker turnover rates, along with a slate of related contributory statistics. First, we calculate the total number of unique workers employed at some point in the year by each MEPS-IC establishment and firm simply by counting the number of W2 records associated with each entity. This employment measure reflects the total number of individuals employed by the given establishment or firm throughout the entire year, including new and departing employees. Employee turnover, as well as company growth or shrinkage, should cause this measure to differ from the existing MEPS-IC employment variables, which measure "typical" or steady-state employment (i.e., the number of workers at a particular point in time). While the ratio of W2-derived total-over-the-year employment and MEPS-IC steady-state employment can be used to approximate employee turnover, we also calculate firm- and establishment-level worker turnover rates following the approach used by the Census Bureau's Quarterly Workforce Indicators (Abowd et al., 2005). Specifically, we use W2s to calculate an establishment or firm's worker turnover as $[0.5 * (\text{hires plus separations}) / \text{steady-state employment}]$. In this context, hires are calculated as the number of workers associated with a firm or establishment that do not have a W2 with the same firm in the prior year, while separations are calculated as the

¹⁰ Note that we do not keep firm-worker matches if none of a firm's MEPS-IC sampled establishments can be matched to their workforces. Conceptually, this occurs when we can match a given firm to a collection of employees, but we cannot find a subset of those employees that plausibly represent the workforce for any of the firm's establishments that appear in the MEPS-IC. For example, this could occur if we match a firm to just 13 workers, but its establishment appearing in the MEPS-IC claims to employ 120 workers—meaning there is no subset of the firm's 13 workers that plausibly represent the establishment's 120. As the example suggests, the cases where all of a firm's MEPS-IC establishments fail to match to a workforce often are a result of the firm-level match itself performing poorly. By dropping firm-worker matches that do not accompany at least one successful establishment-worker match, we ensure that the sample of firms used when presenting firm-level statistics drawn from the MEPS-ICAR is the same used as when presenting establishment-level statistics.

number of workers that do not have a W2 with the same firm in the next year.¹¹ For a W2-derived steady-state employment measure, we calculate the number of workers with W2s at an establishment or firm that have W2s at the same firm in the next year, capturing something akin to December 31st/January 1st point-in-time employment.¹² This turnover measure derived solely from W2 data yields figures that tend to match the approximate turnover measure discussed earlier, in part because the W2-derived steady-state employment measure tends to match the MEPS-IC steady-state employment measure.

A Choice of Employment Concepts: Over the Year vs. Point in Time

An important consideration when analyzing firm/worker- and establishment/worker-level data from the MEPS-ICAR is that the MEPS-ICAR captures all individuals employed by a firm or establishment over the course of an entire year. This is true regardless of whether an individual worked for a MEPS-IC employer for 12 months out of a year or just 12 days. As a result, employee turnover will cause the MEPS-ICAR to link a larger pool of workers to each establishment or firm than are employed by it at any particular point in time. If turnover rates vary by employee characteristics, the characteristics of the over-the-year pool of workers will vary from the point-in-time workforce. For example, if younger workers and low-family-income workers have higher turnover rates on average at a given firm, then the MEPS-ICAR worker pool for that firm will be younger and lower in family income on average than the pool of employees working at the firm at any particular point in time.

While the MEPS-ICAR's default over-the-year employment concept is appropriate for many purposes, there are also circumstances where it may be analytically preferable to present estimates representative of employment just at a given point in time. This is particularly true when seeking to compare MEPS-ICAR data to that from other data sources that adopt a point-in-time employment concept. For example, the Current Population Survey asks workers about their employment situation for particular reference weeks, while the MEPS-IC itself asks establishments about their workforce for a "typical" reference period. In order to facilitate analyses of MEPS-ICAR data using a point-in-time employment concept, the MEPS-ICAR includes a set of point-in-time (PIT) weights that convert MEPS-ICAR estimates from targeting an over-the-year employment concept to a point-in-time employment concept. Conceptually, the weights do this by weighting each

¹¹ For establishments, this method ignores employee turnover generated by movement of workers across establishments within the same firm, and so produces an underestimate of turnover at establishments of multi-establishment firms. Unfortunately, it is effectively not feasible to construct establishment-worker linkages for establishments outside their MEPS-IC survey year, and so calculations that adjust for this type of turnover are not possible.

¹² Strictly speaking, this measure will underestimate steady-state employment when workers quit on December 31st and have replacements start on January 1st. It also might not strictly correspond with end-of-year employment at businesses with strong seasonal employment patterns.

worker matched to an establishment or firm by an estimate of the percentage share of the year during which they worked for that establishment or firm. These weights thus can be thought of as giving the probability that a given worker would be observed if collecting data for a randomly chosen reference day within the year. The resulting weights thus target a point-in-time employment concept akin to measuring an employer's workforce on an average day (as opposed to a specific day, like June 8), not unlike how the MEPS-IC asks surveyed establishments to report on a "typical" work period. The resulting PIT weights also reduce the degree to which workers can influence the data by appearing in the MEPS-ICAR data more than once. The same worker might be employed by several different MEPS-ICAR establishments or firms over the course of a given year, but in PIT-weighted terms, they will not generally be assigned a full year's worth of weight at each job unless they really did work those jobs simultaneously over the full year. In Section IV, we show that after application of PIT weights, MEPS-ICAR estimates of the income distribution and other workforce characteristics tend to be quite close to American Community Survey estimates.

Limitations

The MEPS-ICAR data faces a number of limitations. First, the match between establishments and their workforces is necessarily inexact—a fact that likely injects some measurement error into MEPS-ICAR variables. Second, for estimates using data derived from either IRS Form 1040 data or Decennial Census data, estimates can only be shown for the subset of workers (i.e., W2s) that can be linked to these other data sources. While the linkage between these data sources developed by the Census Bureau is of high quality (Wagner and Layne, 2014), the linkage rate is not 100 percent: the Decennial Census does not collect social security numbers, hampering linkage to W2 data, while not all workers are required to file Form 1040s—a phenomenon that compounds with any other linkage difficulties that may be present. Finally, one key limitation of the MEPS-ICAR dataset is that it does not contain direct information about whether particular workers have health insurance coverage and, if so, whether they obtain this coverage from their employer, which of their employer's plans they are enrolled in, and what type of coverage they have (i.e., single, employee-plus-one, or family). That is to say, while we observe the health insurance choice set that establishments present to their workers, and the number of employees enrolled in each plan and type of coverage, we do not observe the actual choices particular workers make from among the options presented to them. This limitation exists, for the most part, because IRS data on workers' health insurance premiums is not available to us, thus preventing formation of worker-plan links on the basis of cross-referencing those premiums with the premiums MEPS-IC establishments report for their offered insurance plans. In order to proceed, analyses that require linking workers to their choice of plan and type of coverage must simulate that choice using the range of information on

workers' family incomes, the presence of a co-earner in the family, the number of dependents, and other variables available on the MEPS-ICAR.

III. MEPS-ICAR Construction Methodology

The construction of the MEPS-ICAR entailed three main steps: data preparation, linkage of firms to their workforces using identifiers directly available in the Business Register and in W2 records, and the assignment of workers to establishments. For brevity, we do not describe every step in this process, but rather provide an outline of the major assumptions and procedures used.

Stage 1: Prepare the MEPS-IC and Administrative Record Data

In this stage, data from the 2005, 2006, and 2008–2017 MEPS-IC surveys are combined and harmonized. Each year, the MEPS-IC sample is drawn from a preliminary version of the Business Register.¹³ Because information obtained from this preliminary version, including Employer Identification Numbers (EINs; i.e., the identifiers used for firms and subparts of firms within tax records), as well as multi-establishment firm indicators, employment, and annual payroll, can be outdated and can cause tax data linkages to fail, we updated all identifiers and variables using values from the most recently available version of each year's Business Register. This updating process imposes certain consistency safeguards, including rejecting updates to implausible employment and payroll values (e.g., zero employment) which can occur occasionally for various reasons, including data-collection timing issues.

Next, to identify all possible workers employed by a MEPS-IC establishment's parent firm, we extract from the Business Register all EINs associated with firms that contain at least one sampled MEPS-IC establishment in a given year. When doing so, we undertake a range of efforts to ensure the constellations of EINs we associate with firms are internally consistent and do not feature any missing EINs.

The final step in Stage 1 consists of preparing the IRS records and Decennial Census data for eventual linkage with the MEPS-IC. In addition to basic data cleaning and harmonization work, this also involves deduplicating the W2 records, which we do following the recommended practices of McCue and Stinson (2019). At this stage, we also link the W2s with Form 1040s and Decennial Census data from 2000 and 2010. This linkage is fairly straightforward, as all of these datasets share a common person-level identifier previously constructed by the Census Bureau (Wagner and Layne, 2014).¹⁴

¹³ For more details about the Business Register's construction, see DeSalvo, Limehouse, and Klimek (2016).

¹⁴ This identifier is known as a Protected Identification Key (PIK) and was constructed by the Census Bureau to protect personally identifiable information (PII).

Stage 2: Produce the Firm-Level Link and Prepare for Forming Establishment-Level Links

Using the datasets prepared in Stage 1, we link workers to MEPS-IC firms by matching the EINs listed on workers' W2s to all EINs associated with MEPS-IC establishments and their parent firms.¹⁵ This fairly straightforward matching process is sufficient to link MEPS-IC firms to all of their employees. However, for firms with more than one establishment, it does not complete a match between workers and the specific establishments at which they work, principally because firms may file taxes using the same EIN for more than one establishment. In Stage 3, we constructed worker-establishment matches for employees of multi-establishment firms, though to do that we first needed to construct several auxiliary data sets.

The first key auxiliary data input is geocoordinates for the residential address of each worker. In most cases, we obtain residential addresses for workers from their Form 1040s. However, if unavailable, location information is derived from the temporally nearest of the 2012–2017 Resident Candidate File, 2010 Decennial Census, and 2000 Decennial Census data. When geocoordinates for exact addresses are unavailable from these sources, we assign workers to the population-weighted centroid of their residential address's zip code. When no location information is available beyond their state of employment, which is always available from workers' W2s, we assign workers to an imputed set of geocoordinates based on the location of other employees in the same firm that work in the same state.

The next key set of auxiliary data consists of employment targets for MEPS-IC establishments that are part of multi-establishment firms. These targets represent the number of distinct W2s that we should expect to find among the W2s linked to a given establishment. A starting point for these targets is the employment and payroll levels reported by establishments in the MEPS-IC and listed in the Business Register. However, since MEPS-IC employment figures reflect "typical" levels of employment at any given time, rather than the sum of all individuals who worked at the establishment during the calendar year, these MEPS-IC employment numbers will generally be smaller than the number of W2s that should link to the establishment. For example, an establishment might report a typical employment level of 10 workers, but would have 15 W2s if, over the course of the year, it had 5 workers quit and hired 5 new workers to replace them. Note that these targets are only needed for establishments that are part of multi-establishment firms, since the

¹⁵ The MEPS-IC is a survey of establishments, and some multi-unit firms may have one or more but not necessarily all of their establishments in the MEPS-IC sample.

firm-level match also solves the establishment-level match when a firm has only one establishment.¹⁶

To develop these employment targets, we use data from single-establishment firms, including a rich set of predictor variables and the number of W2s to which these establishments match, to train a Least Absolute Shrinkage and Selection Operator (LASSO) model to predict the ratio of W2s to MEPS-IC reported employment (i.e., the number of unique W2s matched to each establishment divided by the total number of employees reported by the MEPS-IC establishment). We then apply this ratio to the MEPS-IC employment total to construct the target employment totals (i.e., target W2 counts) for all multi-establishment firms represented by MEPS-IC establishments. While doing this, we also make additional efforts to ensure that establishments' designations as part of either single-establishment or multi-establishment firms are consistent with all available data and that W2-to-establishment-employment ratios from mislabeled establishments are not used to train the LASSO model.

In addition to constructing employment targets, we also build target total payroll figures for establishments. These reflect the total amount of payroll we expect to result from summing W2 pay across all workers matched to an establishment in a given calendar year. In general, there is less need for adjustment when moving from establishments' Business Register-derived annual payroll totals to the corresponding quantities in the linked W2 data, as there is no "typical" period vs. "calendar year total" mismatch for payroll figures in the way that there is for employment figures. Therefore, our procedure here consists of a fairly simple two-step process. First, for each establishment, we calculate an adjustment ratio that consists of all W2 payroll linked to its parent firm divided by its parent firm's Business Register-derived annual payroll total. Second, for each establishment, we assign it a target consisting of its own Business Register annual payroll total multiplied by its firm's adjustment ratio, cleaning the ratios and final targets both to censor extreme values and prevent implausible average annual employee wage levels from appearing.

Finally, to simplify the matching process in the next stage, we temporarily consolidate establishments located within 2 miles of one another into single synthetic establishments (summing together their targeted employment and payroll totals). We treat these synthetic establishments as one large establishment in all future steps, until we split them back into their individual components. We also divide multi-establishment firms, where possible, into separate synthetic firms. We do this by creating clusters of establishments (i.e., synthetic firms) defined such

¹⁶ Strictly speaking, we do produce predicted employment targets for establishments in single-establishment firms as well. We use these targets as one of many inputs in our quality assurance exercises that are intended to double check whether these establishments really are part of single-establishment firms and whether their matched pool of W2s is reasonable.

that no establishment in any given synthetic firm is within 400 miles of an establishment from the same actual firm that is placed into a different synthetic firm. In practice, this allows us to treat groups of geographically distant establishments within multi-establishment firms as independent from each other. We apply similar rules to divide employees of such firms into separate synthetic firms.

Below, we generally use the term “establishment” and “firm” to refer to the synthetic establishments or synthetic firms created above. We do this for the sake of brevity and because the procedures in Stage 3 do not distinguish between the synthetic and non-synthetic cases, except when explicitly noted. Later, in Stage 4, synthetic establishments are broken back out into real establishments and synthetic firms are reconsolidated into real firms.

Stage 3: Match Workers in MEPS-IC Firms to MEPS-IC Establishments

At the start of Stage 3, we have the following information for each MEPS-IC establishment in each survey year: a pool of workers (W2s) associated with the firm owning that establishment, a target number of workers to assign to each establishment from their firm’s broader pool of workers, and a target quantity of total W2 payroll to find for each establishment in the MEPS-IC. To match workers to each MEPS-IC establishment, we proceed as follows. First, for each MEPS-IC firm in each year, we check how many of its establishments we observe in the MEPS-IC sample that year. The matching approach differs across each of the following three types of cases that we observe:

- Case 1: one establishment in the MEPS-IC, drawn from a parent firm that has no other establishments.
- Case 2: one establishment in the MEPS-IC, drawn from a parent firm that has additional establishments.
- Case 3: multiple establishments in the MEPS-IC that share a parent firm.

Case 1: Single-Establishment Firm

For a given establishment, if we flag it as the only establishment in its firm and have verified that this status is consistent with the observed tax data, then the firm-level match to W2s has already solved this establishment’s match and no further work is required.

Case 2: One Establishment from a Multi-Establishment Firm

In Case 2, where the MEPS-IC samples only one establishment from a firm with multiple other establishments, we proceed as follows. To begin, we assess the feasibility of achieving an assignment of workers to the establishment that achieves both employment and payroll totals within a tolerance range of the targeted

values.¹⁷ Our test for feasibility is quite permissive, and it contains two parts. First, supposing that X is the least number of workers that can acceptably be assigned to the establishment (i.e., supposing that X is the lower bound of the establishment's employment tolerance range), we check whether assigning the establishment the X lowest paid workers in the firm would yield an assignment with an unacceptably high amount of payroll. Second, supposing that Y is the greatest number of workers that can acceptably be assigned to the establishment, we check whether assigning the establishment the Y highest paid workers in its firm yields an assignment with an unacceptably low amount of payroll. We consider there to be no feasible assignment for an establishment if either of these tests fail.

Next Steps for Case 2 if a Match is Feasible

If it is feasible to find an assignment of workers that meets both the employment and payroll targets for the establishment, we proceed by calculating the distance from each worker to the establishment and array workers in order from closest to furthest from the establishment. We then apply the following algorithm:

In Step 1, we check whether there is a number of workers N such that (a) N falls within a tolerance range around our target number of workers, and (b) summing workers' total W2 pay from worker 1 (the worker closest to the establishment) to N yields a payroll total within a tolerance range of the total payroll target. If such an N exists, we assign the N closest workers to the establishment and consider the match complete. When multiple Ns satisfy these conditions, we select the N that minimizes the sum of squared differences between actual totals and the employment and payroll targets (dividing all payroll figures by 60,000 prior to computing squared differences).¹⁸

If no such N exists, in Step 2 we assess why this is the case. If the problem is that all assignments featuring an acceptably large amount of employment within the current ordering of workers assign too much payroll to the establishment, then some relatively high-pay workers need to be eliminated from consideration for assignment and replaced by relatively low-pay workers. To do so, we calculate the average level of pay among the closest workers too far from the establishment to be initially assigned to it. We then calculate the average pay level among each of the 10 percent and 10 to 25 percent highest paid workers that are close enough to

¹⁷ For employment, this tolerance range is generally equal to the targeted employment value plus or minus the greater of 3 workers and 20 percent of the target value. For payroll, this tolerance range is generally the targeted payroll total plus or minus the greater of 25 percent of the targeted level or \$50,000.

¹⁸ We divide payroll figures by 60,000 so that employment and payroll squared differences are in more comparable units: without this adjustment, a discrepancy of \$5,000 in total payroll between an assignment of workers and an establishment's payroll target would be weighted equally to a discrepancy of 5,000 workers from an establishment's employment target. The payroll scaling factor 60,000 is a bit high in the sense that the mean MEPS-IC job is paid just under half of \$60,000 (see Table 7), but it was chosen so as to weight the match slightly in favor of fidelity to employment targets over fidelity to payroll targets.

have been provisionally assigned to the establishment. Using these figures, we calculate a number of workers from these two pay-level ranges to eject from the commute distance ordering that will, on average, result in a new ordering that contains an acceptably sized assignment of workers with total payroll that is either within or is as close as possible to being within the targeted payroll tolerance range. We then eject the calculated number of workers from consideration for assignment, selecting the specific workers to eject from each pay level range at random. Once this ejection process is completed, we return to Step 1, testing whether or not there is a target-satisfying assignment of workers within the new arrangement. We handle the case where too little payroll is assigned to the establishment symmetrically. We then iterate between Steps 1 and 2 until an assignment is found or until it becomes infeasible to find an assignment that satisfies both the employment and payroll targets among workers that have not been ejected. If we enter this latter case, we return all ejected workers to the candidacy pool and return to Step 1. If no solution results after a large number of iterations through this procedure, we revert to random assignment of workers from among a set of workers relatively close to the establishment (i.e., from among the $2.5 * \textit{employment target}$ closest workers).

Next Steps for Case 2 if No Match is Feasible

If we find that there is no collection of workers from the full set of candidates that satisfies the establishment's employment and payroll targets, a different assignment strategy is pursued, depending on why the targets could not be met. If no feasible match exists because there are both too few workers and too little payroll, we assign all workers to the sampled establishment and complete the match, bearing in mind that we may reject this match later for failing to meet quality standards in Stage 4. If there are enough workers to achieve an assignment within range of the employment target but the smallest number of workers we can assign still brings too much payroll to the establishment, we proceed with one of two approaches. If deviation from the target range exceeds the upper end of the target range by a large factor (i.e., it falls outside two times the tolerance range), we assign the establishment its geographically nearest employees until it achieves a within-target-range employment level. If deviation from the target payroll range is not too large, we assign the establishment the lowest paid workers from among a set of workers relatively close to the establishment in terms of commute distance until a quantity of employment within the target range is achieved.¹⁹ Symmetric procedures using the highest paid workers are applied when the problem with feasibility is inability to assign enough payroll to the establishment.

Case 3: Multiple Establishments from a Multi-Establishment Firm

In this case, we must assign workers to multiple establishments from the same firm and thus from the same pool of workers. Here, we begin by calculating the distance

¹⁹ Specifically, we assign the lowest paid workers from among the $5 * \textit{employment target}$ closest workers to the establishment.

from each worker to each establishment. We then form a provisional assignment of workers to establishments by assigning to each establishment a number of workers as close as possible to its employment target from among those workers that are closer to that establishment than any of the other establishments in the same firm. We then loop through establishments several times, adding workers when the given establishment does not have enough and, where possible, replacing workers assigned to the given establishment with closer workers not assigned to any other establishment. The purpose of this provisional assignment is just to give each establishment a starting set of workers of an appropriate number, with some effort to control commute distances.

Once these provisional assignments have been made, we begin a process of allowing the establishments within the same firm to trade workers with one another. We cycle through establishments, permitting establishments to do each of the following actions once per cycle: trade an employee with another establishment, trade an employee with the unassigned employee pool, donate an employee to the unassigned employee pool, and take an employee from the unassigned pool. In each cycle, the exact pair of workers traded between two establishments is chosen at random among the set of trades that moves both establishments closer to their employment and payroll targets without causing either establishment to add a worker with an overly long commute distance. Similar restrictions apply to an establishment seeking to take an action with the unassigned pool, though no restriction is made on what happens to total payroll and employment within the unassigned pool. Once all establishments are assigned a set of workers that meet their employment and payroll targets, or once a very large number of trades have been completed, the trading process stops, and the assignment is finalized. As the maximum trade limit approaches, we adjust the worker-trading-pairs selection process to make increasingly aggressive trades that are more tolerant of disadvantageous effects on commute distance.

Stage 4: Finalize the Match and Identify Failed Matches

After completion of Stage 3, all establishments (and synthetic establishments) have a set of assigned workers. However, some additional processing is required before the data can be finalized. The most straightforward component of this work consists of dropping the synthetic firm labeling and switching back to labeling establishments in accordance with their actual parent firms.

Additionally, synthetic establishments must be split back into their constituent actual establishments. We do this by randomly assigning workers from the synthetic establishment's worker pool to its constituent establishments, in proportion to each actual establishment's share of the synthetic establishment's employment. Then, we allow the actual establishments to go through trading cycles to improve their assignments' proximity to their payroll targets, in a fashion

analogous to those for Case 3 trades in Stage 3 between establishments within multi-establishment firms. The trades here differ from those in Case 3 mainly in that (a) we ignore commute distance, since all constituent establishments of a synthetic establishment are necessarily geographically very close to one another; (b) workers are not permitted to enter an unassigned worker pool, meaning all trades must be between establishments; and (c) we allow a limited number of worker donations from one establishment to another. As before, the trading cycles are complete once all actual establishments have employment and payroll totals within a tolerance range of their target values or after a very large number of trades have been completed.

Once we have split synthetic establishments back into actual establishments, we are quite close to having a finalized link between MEPS-IC establishments and their workforces. The final step consists of identifying failed matches. We define a match as having failed when the number of expected workers matched to an establishment or firm deviates very severely from the number actually linked. The most clear-cut case of match failure is when no workers can be found in the W2 data for a given firm and all of its establishments. A case where, for example, only 1 worker is found where 100 are expected would also trigger a match failure. The precise thresholds for match failure depend on the size of the establishment, but in general are calibrated to preserve as many matches as is reasonably possible and thereby tolerate considerable variation across firms. Match failures often arise from the algorithms specified in Stage 3 when the available worker pool is too small for employment and payroll targets to be achieved, suggesting problems with the firm-worker match. Therefore, when match failures occur, we delete all linkages (establishment and firm) associated with the failed match. Note that the (successful) match rate is available in Table 1.

Stage 5: Produce Turnover Statistics and Point-in-Time Weights

In this final stage of data production for the MEPS-ICAR, we begin with a finalized match between MEPS-IC firms, MEPS-IC establishments, and their workforces with all match failures removed. In this stage, we complete a range of largely anodyne data cleaning and variable construction tasks for the convenience of final data users. We also create the establishment-level and firm-level roll-up files that offer establishment- and firm-level summary statistics of the worker-level data.

Next, we produce a set of employee turnover and steady-state employment statistics, defining turnover following the Census Bureau's Quarterly Workforce Indicators (QWI) definition of $[0.5 * (\text{hires} + \text{separations}) / (\text{employment})]$ (Abowd et al., 2005). We calculate turnover statistics at the firm level by examining the set of workers linked to a firm in each year, calculating the firm's hires for the year as the number of those workers that did not have a W2 associated with that firm in the prior year and calculating the firm's separations for the year as the number of

those workers that did not have a W2 associated with that firm the following year. We take the firm's steady-state employment to be the number of workers at the firm that did have a W2 associated with it in the following year. This approach suffices to give us turnover measures in any year where we can access W2 data in the surrounding years. When we only have one year of neighboring W2 data,²⁰ we calculate steady-state employment relative to whichever year we have and then replace the (hires + separations) component of the formula with 2 times whichever data element we do observe. We take a similar approach to calculating establishment-level turnover, with the proviso that we add an assumption that establishments never gain or lose workers to other establishments within the same firm.

Finally, we produce a set of point-in-time weights that contain an estimate of the percentage share of the year each worker was employed by their matched establishment or firm. We produce these weights as follows, generating one set of weights for firm-level analyses and another for establishment-level analyses. We begin by creating an initial set of candidate weights that sums across matched workers to each establishment or firm's steady-state employment level. These weights assign workers an initial weight of 1 (i.e., a weight representing year-round employment) if they appear to have been employed by the same firm in the years surrounding their MEPS-ICAR reference year.²¹ All other workers are assigned a lower weight equal to the establishment's steady-state employment level less the number of workers assigned an initial weight of 1, all divided by the number of workers not assigned an initial weight of 1. We then adjust these initial weights based on a number of assumptions. In particular, we assume that workers with very high incomes worked year-round for their employer and that workers with very low incomes did not work for their employer for more time than it would take to earn their pay if they worked 15 hours a week at the minimum wage in their year of employment. We further assume that workers observed at the start (or finish) of a multi-year-long job spell worked at their employer for a portion of the year in their first (or last) year of employment equal to their first (or last) year salary divided by their next (or prior) year's salary, with an inflation and income growth adjustment. We finish weight production by adjusting the modified weights until point-in-time weighted employment for each establishment or firm once again matches the establishment or firm's steady-state employment. We do this by shrinking the individual weights toward the average weight that would sum to the correct steady-state level. We also supplement these primary point-in-time weights with some ancillary ones targeting beginning-of-year and end-of-year point-in-time employment. We produce the beginning-of-year weights by assigning a weight of 1

²⁰ We currently have W2 and Form 1040 data only for the same years the MEPS-ICAR has MEPS-IC data: 2005, 2006, and 2008–2017. We plan to expand the MEPS-ICAR as more years of input data become available to us.

²¹ Similar to when we calculate turnover statistics, data constraints require that we ignore the possibility of employee transfers between establishments within the same firm for the purposes of these calculations.

to all workers that could be matched to an employment record from the same firm in the year prior to their reference year, and a weight of 0 to all other workers. For end-of-year employment, we produce the weight similarly, but focusing on workers that can be matched to an employment record from the ensuing year. Our recommended point-in-time weights, however, are those that use the fuller suite of adjustments described above.

At this stage, we have completed construction of all components of the MEPS-ICAR dataset.

IV. Assessing the MEPS-ICAR

Establishment- and Firm-Level Statistics

In this section, we present assorted statistics calculated at the establishment and firm levels intended to characterize the quality of the MEPS-ICAR. We begin by considering match rates between establishments and their workforces in Table 1. In Tables 2A, 2B, 3A and 3B, the focus is on how well the distribution of employment, employee turnover, and payroll in the MEPS-ICAR matches expectations. In Tables 4A and 4B, we re-examine the employment, turnover, and payroll statistics among establishments of single- and multi-establishment firms separately, doing so because of differences in the matching algorithm used between these cases. We conclude our review of establishment- and firm-level statistics by considering a set of quality-test regressions in Table 5, before moving on to worker-level statistics.

Match Rates

We begin by examining the rate at which establishments successfully match to their workforces. The first three columns of Table 1 show successful workforce match rates for MEPS-IC establishments overall, for establishments that offer health insurance, and for establishments that do not offer health insurance. The next three present those same match rates, but with employment weights (i.e., MEPS-IC–reported establishment employment multiplied by the establishment survey weight). Match rates of these sorts are also presented in this table by year, Census division, industry, single- vs. multi-establishment firm status, for-profit versus non-profit status, firm size category, and establishment size category. All match rates also include statistical significance indicators comparing the match rate in the specified category against all other establishments. It is worth bearing in mind, however, that all but the smallest of differences tend to be statistically significant when comparing very broad national samples pooled across multiple years.

Table 1 indicates that the overall establishment-level successful match rate is 92.89 percent, with the match rate being somewhat higher among establishments that offer health insurance (94.00 percent) than among those that do not (91.73 percent). In employment-weighted terms, the overall match rate is 93.33 percent,

with little heterogeneity between establishments that do and do not offer health insurance. The match rates by year point to lower establishment match rates for the years 2008–2011, with this reduced match rate being driven principally by a reduced match rate among establishments that do not offer health insurance. We speculate that this may be related to survey data quality issues caused by the Great Recession.

Match rates are quite similar across Census divisions, with no Census division's match rate falling considerably outside the 91–95 percent range. Match rates by industry are similarly clustered, with the exception of the employment-weighted match rate for the Agriculture, Fishing, and Forestry sector, which dips to 88.81 percent overall. Establishments in single-establishment firms and multi-establishment firms overall have fairly similar match rates, though establishments in single-establishment firms tend to have match rates that are a few percentage points higher in employment-weighted terms. For-profit and non-profit establishments also have fairly similar match rates, with non-profits generally having match rates that are a few percentage points higher than for-profit establishments. Finally, the match rates by firm size and establishment size show that single employee firms tend to have the lowest match rates of all, with an overall match rate of 87.80 percent. Similar match rates are found for all other sizes of establishments and firms and fall in the 92–95 percent range. Overall, spanning across the entire time period, we successfully match approximately 328,000 MEPS-IC establishments and 280,000 MEPS-IC firms, failing to match only 26,000 establishments and 23,000 firms total.

Employment and Turnover

Next, Tables 2A and 2B explore the data on employment and turnover, presenting means and assorted percentiles in different panels for MEPS-IC employment- and establishment-weighted estimates.

The first two rows of results show the distributions of the targeted number of workers to be found for each establishment (as derived in Stage 2 of the matching process described in Section III of this paper) and the number of workers actually matched. The mean establishment had a worker count target of 29.84, with 27.37 having actually been matched. The percentile estimates point toward fairly close matches to the targets, with the number matched generally falling behind only at larger establishments. This is borne out by the employment-weighted version of these statistics, which point to the mean worker target being about 1,155 workers, with the mean number of workers matched being 965.3. The question of how well we match targets takes for granted that the targets are appropriate. One external check on that assumption comes in the next two rows, where we give the actual reported establishment employment levels in the MEPS-IC in one row, and in the next, the level of W2-derived steady-state employment (i.e., the employment measure used when producing our turnover statistics), which we can construct only

after the match has been completed. To the extent these figures match, they suggest that the matched sample of workers has properties implying similar steady-state employment levels to those reported by MEPS-IC respondents. Here, we see that the mean establishment reports a steady-state employment of 16.96 workers, while our turnover statistics imply a quite similar level of 18.74 workers, suggesting that the MEPS-ICAR matches are generally of high quality.

In the next block of Tables 2A and 2B, we examine ratios of matched worker counts to target worker counts, ratios of matched worker counts to MEPS-IC reported employment, and our estimate of turnover. The ratio of matched to target worker counts gives a natural means of assessing how well the match procedure hit its targets. The mean and median values of this ratio are 1.05 and 1.00 respectively, or 0.93 and 0.93 respectively in employment-weighted terms. In the tails of the establishment-weighted distribution, ratios quite discrepant from 1 are possible. Ratios where the number of matched workers is large relative to the target tend to exist mainly due to very small establishments (e.g., finding two workers when one is targeted), and these large ratios tend to be muted in employment-weighted terms. Ratios in the neighborhood of 0.6 are more prevalent in the low-end tails even with employment weights, however. This undershooting tends to occur at least in part because of cases where the number of workers available in the worker pool was small relative to the number of workers anticipated for matching.

After looking at the matched worker count to target worker count ratio, we then examine the ratio of matched worker counts to MEPS-IC reported employment in relation to our formal estimate of turnover, bearing in mind that the former should (barring substantial over-the-year changes in firm size) approximately equal the formal turnover estimate plus one. Here, we see that the mean establishment has a turnover rate in the approximate sense (i.e., the matched worker count to MEPS employment ratio) of 64.4 percent and the median establishment has a turnover rate of 37.8 percent. The approximate turnover rate for many establishments is 0. In the extreme tails, the approximate turnover rate can be negative or can exceed 238 percent. While the negative values are necessarily spurious, large turnover rates are not necessarily inappropriate, as there are businesses where very high turnover rates are common. Also note that the negative approximate turnover rates are muted once employment weights are applied, even as the mean and median values do not substantially change. Comparing these ratios to the formal turnover measure derived from the IRS data alone, we see that the formal turnover rate is 46.3 percent and 26.1 percent for the mean and median establishments respectively, or 52.2 percent and 34.0 percent at the mean and median respectively when employment weights are applied. These suggest that the IRS turnover rates are systematically a bit lower than those implied from comparing MEPS-IC establishments' reported employment levels to their number of matched workers.

Having assessed some establishment-level employment match statistics, we next consider some statistics on the quality of the match in employment terms at the firm level. We present the same statistics as for establishments, except without any figures relevant for worker targets since no targets are formed at the firm level. First, looking at raw employment totals, we find that MEPS-IC establishments tend to report firm-wide employment levels that are reasonably close to, though larger than, the steady-state employment levels implied by our turnover statistics. Second, we find that the turnover rates implied by the ratio of the number of W2 workers matched to a given firm to MEPS-IC reports of firm employment are, at the mean, considerably larger than the turnover rates implied by what we can observe in the IRS data, though the two measures are quite close at the median. Our view is that this does not necessarily reflect poor-quality firm-worker matches in the MEPS-ICAR, so much as that there is a long right tail of establishments that report severe underestimates of employment at their parent firm in the MEPS-IC survey.

Payroll

Tables 3A and 3B examine the match in terms of payroll. We find that the targeted and matched establishment-level payroll totals tend to be fairly similar. The mean establishment has 23.9 percent more matched payroll than its target would suggest, though the 25th- through 75th-percentile establishments have exactly the anticipated amounts. Over 90 percent of establishments are matched to a quantity of payroll within 80 percent of the targeted level. Application of employment weights implies that the mean worker works at an establishment matched to 2.6 percent more payroll than its target would suggest—a considerable improvement in accuracy relative to the baseline without employment weights. Statistics comparing matched payroll totals to raw MEPS-IC/Business Register establishment-level payroll totals are also provided. While the matched payroll totals are similar to the Business Register totals across most of the distribution, the mean ratio of matched payroll to Business Register payroll is 11.21. This very large ratio results from the presence of a small number of cases where Business Register totals are dramatically lower than matched payroll totals. These outlier ratios appear to be generated by cases where the Business Register's firm-level payroll total variable instead reports an establishment-level payroll total.²² The presence of problems associated mainly with extreme outliers in this context highlights the importance of our use in the match of lightly edited payroll targets that do not have this problem.

Tables 3A and 3B also presents information on how well the matched payroll totals correspond with MEPS-IC/Business Register payroll totals at the firm level. The mean and median firm have matched and Business Register payroll totals within less than 3 percent of one another. In employment-weighted terms, the matched total is about 10 percent lower than the Business Register payroll total, though the median value is within 1 percent of the Business Register total. Since no targets are

²² This may also be attributed to not all EIN units reporting, or timing issues related to a restructuring of the organization.

produced at the firm level, these comparisons do not include adjustments for outlier Business Register payroll reports, though severe outliers of any type tend to be considerably less common at the firm level than at the establishment level in the Business Register.

Employment and Payroll by Single-Establishment vs. Multi-Establishment Firm Status

Matching establishments to their workforces is considerably more difficult within multi-establishment firms than in single-establishment firms. To check on match quality for establishments that are members of these two different types of firms, Tables 4A and 4B present a simplified set of the establishment-level employment and payroll statistics from Tables 2A, 2B, 3A, and 3B, with Table 4A showing data for single-establishment firms and Table 4B displaying data for multi-establishment firms. The figures in Table 4B do not suggest significant degradation of match quality when matching establishments within multi-establishment firms. The mean establishment in a single-establishment firm has a matched worker count that exceeds its target by about 7.2 percent; the mean employee of a single-establishment firm works in an establishment matched to about 6.1 percent fewer workers than its target would suggest. The same figures for multi-establishment firms are 0.4 percent and 7.1 percent respectively. Median match fidelity to target is arguably better at establishments of multi-establishment firms. Establishments of both types of firms generally also had post-match W2-derived employment levels that matched their MEPS-IC employment levels fairly closely, as well as estimated turnover rates that corresponded to their matched workers to MEPS-IC employment ratios within a reasonable tolerance.

The results above offer little cause for concern about establishment matches for the multi-establishment firm case relative to the single-establishment firm case, at least in terms of employment. However, this in part reflects the fact that the matching algorithm used tends to prioritize hitting employment targets over payroll targets. A more complete consideration of the match requires checking on payroll target performance as well. Table 4A indicates that the single-establishment-firm match generally gives the mean establishment about 15.9 percent too much payroll relative to target, or about 6.7 percent too much at the establishment employing the mean employee of a single-establishment firm. Performance of the match for establishments in multi-establishment firms does tend to degrade somewhat. In particular, matched payroll totals tend to exceed targeted totals by 44.8 percent at the mean multi-establishment-firm establishment. However, in employment-weighted terms, the mean difference is only 0.2 percent, suggesting that the divergence between matched and targeted totals is driven mainly by large proportional discrepancies at small establishments that might not be particularly large in absolute terms. Finally, note that the figures comparing matched payroll totals to Business Register totals exhibit the same problem with extreme mismatch at the mean for establishments of multi-establishment firms as do the overall

numbers. As discussed above, this mismatch at the mean is driven by a small number of very extreme outliers among multi-establishment firm establishments where Business Register firm-level payroll totals appear to actually be reporting establishment-level totals. Other than this issue with outliers affecting the mean, which data cleaning efforts eliminated from the payroll targets we actually use in the match, the ratio of matched to Business Register payroll tends to be quite similar to the matched to target payroll ratio across most of the distribution for both subsets of establishments. Overall, mean and median performance of the match in payroll terms seems quite good for both single-establishment and multi-establishment firms and comparable in quality to what is suggested by the employment and turnover data.

In addition to the above general tests of match performance, we also investigated whether match performance in terms of fidelity to employment and payroll targets varies considerably by year, industry, Census division, establishment and firm size category, and whether or not an establishment offers health insurance. In results available upon request, we find very limited qualitative variation along these dimensions. The only exceptions are that we are more likely to overshoot employment and payroll targets in the Agriculture, Fishing, and Forestry sector as well as at establishments with five or fewer employees.

Alternative Metrics on Match Quality

In Table 5, we present an alternate approach to considering match quality. Here, we present results of simple univariate regressions of MEPS-IC and Business Register variables on their matched equivalents. Namely, using establishment-level variables, we regress MEPS-IC employment on number of matched workers, the employment targets on number of matched workers, Business Register payroll on matched payroll, target payroll on matched payroll, the share of workers that are women reported on the MEPS-IC versus the same share among matched workers, and the share of workers aged 50+ per the MEPS-IC versus the same figure among matched workers. We also estimate regressions at the firm level for MEPS-IC employment versus number of matched workers and Business Register payroll versus matched payroll. We run these univariate regressions in the first panel using MEPS-IC survey weights, thereby obtaining establishment-weighted estimates, and with MEPS-IC survey weights multiplied by employment totals in the second panel, thereby obtaining employment-weighted estimates. Quality match performance should generally be indicated by high R-square values and regression coefficients relatively close to 1, except when regressing (other than target) employment variables on matched worker counts.

At the firm level, Table 5 yields R-squares exceeding 90 percent for both the employment and payroll regressions, with or without employment weights. The payroll coefficients are also generally close to 1. At the establishment level, the regressions using employment and payroll, regardless of choice of weights,

generally have R-square values in the 80–90 percent range, with the exception of the establishment-weighted target workers regression (R-square of 91.7 percent) and the establishment-weighted Business Register payroll regression (R-square of 66.9 percent). The poor performance of the establishment-level Business Register payroll totals here matches with what was observed in the prior summary statistics. In the regressions checking the demographic statistics, the percent-women regressions generally had performance comparable in terms of R-squares to the employment and payroll regressions, though the R-squares for the percent-aged-over-50 regressions were in the 70 percent range. Overall, we would characterize these regressions as qualitatively favorable signs for the quality of our match.

Worker-Level Statistics

In this section, we consider a set of statistics at the matched-worker level. For all statistics presented in this section derived from the MEPS-ICAR, we present them using one of two sets of weights. The first set of weights are just the standard MEPS-IC survey weights, producing estimates representative of the default MEPS-ICAR over-the-year employment concept. The second set of weights also apply our point-in-time (PIT) weights, producing estimates representative of a typical point-in-time employment concept. The PIT-weighted estimates should be conceptually more comparable to those from the external survey data sources we will compare MEPS-ICAR estimates against in this section.

Table 6 presents means of certain key demographic variables for the workers matched in the MEPS-ICAR and compares them against comparable figures, where possible, from pooled American Community Survey (ACS) data from the same time period (Ruggles et al., 2021; U.S. Census Bureau, 2005–2017).²³ Table 7 is similar, but presents means plus an additional slate of percentiles for age, family income, and personal income.²⁴ The pooled ACS data that we use includes all individuals in the labor force that have had a job at some point, that do not work in the public sector, that are not in the armed forces, and that do not report having been continuously unemployed for 5 or more years. The ACS comparison pool is set to include all workers in the labor force that are not long-term unemployed, not just those employed at the time of survey, since this is more comparable to the MEPS-ICAR data's workforce concept.

Worker-Level Matches and Demographic Characteristics

Table 6 begins by highlighting the match rate between MEPS-ICAR workers and other data sets, showing that 8.39 percent of MEPS-ICAR workers cannot be associated with a Form 1040, while 6.36 percent of MEPS-ICAR workers cannot be

²³ We obtain our copy of the American Community Survey data from the University of Minnesota's IPUMS USA project (Ruggles et al., 2021).

²⁴ We top code ACS income variables at \$250,000 for improved comparability with the MEPS-ICAR data, whose inputs are also subject to top coding.

linked to a Decennial Census record. Both match failure rates are lower in PIT-weighted terms, falling respectively to 6.28 percent and 5.17 percent. These match failure rates are worth noting immediately, as all MEPS-ICAR worker-level means in this table are presented only within the subset of workers that can be matched to the linked data source (for most variables, this is the Decennial Census, but for the children counts and the marital status variables, this is the IRS 1040 data); some amount of difference between the MEPS-ICAR and ACS data should be expected due to these linkage issues.

Moving to Table 6's demographic estimates, the means for workers' sex and age variables point to the average worker in the MEPS-ICAR dataset, using the default over-the-year employment concept, being about 3 years younger and 2 percentage points more likely to be female than the mean labor force participant in the ACS. Application of point-in-time weights, however, brings the MEPS-ICAR and ACS age estimates within 1 year of one another, though there is little effect of PIT weighting on the female share of the MEPS-ICAR workforce. Next, using the default weights, MEPS-ICAR workers are about 2 percentage points more likely to be non-Hispanic White or non-Hispanic Black than in the ACS. Application of PIT weights does adjust the racial and ethnic composition of the MEPS-ICAR sample, but with little net impact on the degree of correspondence to the ACS data.

Table 6 concludes by presenting means of marital status and family composition variables. Prior to application of point-in-time weights, the MEPS-ICAR has 10 percentage points fewer married workers than the ACS, having instead about 5 percentage points more single workers without children and 5 percentage points more single workers with children. This considerable gap largely closes after applying PIT weights to the MEPS-ICAR estimates. Doing so brings the MEPS-ICAR single workers without children estimate to within 1 percentage point of the ACS estimate. The PIT-weighted MEPS-ICAR married workers estimate is still 5 percentage points lower than the ACS estimate, with the single workers with children estimate being 4 percentage points higher. This remaining wedge may be due in part to a difference in measurement concepts between the two datasets. The ACS data here literally refers to unmarried individuals with children, while the MEPS-ICAR data actually refers to individuals filing their taxes with "Head of Household" status. While this filing status is used by single or unmarried workers with dependents, it can also be claimed by individuals with dependents who are married but separated or married to a nonresident alien. The final estimates in Table 6 show that the typical worker in the MEPS-ICAR has on average 0.074 fewer children at home than the typical ACS worker, with this gap falling to 0.055 fewer children after PIT weighting.

Overall, the differences between the MEPS-ICAR and the ACS in terms of demographic composition and family structure are quite small when one uses point-in-time weights to ensure that conceptually comparable estimates are being

compared. There remain some gaps between the two data sources, especially in terms of family structure. These gaps likely reflect a mixture of differences in underlying measurement concepts, differences in how the ACS defines a family relative to how the IRS defines a tax filing unit, and an imperfect linkage between MEPS-ICAR workers and both IRS Form 1040 and Decennial Census data. Data users should bear these issues in mind when seeking to compare MEPS-ICAR estimates to those from the ACS and other data sources.

Full Distributions of Worker Ages, Wages, and Family Incomes

Turning to Table 7, we can see that worker ages in the MEPS-ICAR seem to be a few years lower than in the ACS across the full age distribution, with use of point-in-time weights largely closing the gap between the two datasets. Next, we look at means and percentiles of the W2 wage income associated with workers' jobs at MEPS-ICAR employers alongside means and percentiles of the ACS sample's reported wage and salary income. The mean over-the-year worker in the MEPS-ICAR is receiving approximately \$9,000 less from their MEPS-ICAR job than the mean worker in the ACS reports receiving in terms of annual wage and salary income. This is to be expected, since the pay MEPS-ICAR workers receive from their jobs will often be pay for jobs that they did not work for the entire year, whereas the ACS report includes pay from all jobs worked over the year (as well as from second jobs held simultaneously). When we apply point-in-time weights to the personal wage income estimates from the MEPS-ICAR, thereby weighting MEPS-ICAR jobs by the share of the year they were actually worked, the gap in personal wage income between the MEPS-ICAR and the ACS closes almost completely across the entire wage distribution, with the means falling within \$2,000 of one another.

Next, we consider the means and percentiles of the distributions of total family income in the ACS versus family total money income reported on Form 1040s. These results show that the mean Form 1040 family total money income in the MEPS-ICAR is \$66,790 at baseline and \$77,710 after PIT weighting, while the ACS total family income is \$79,526. The PIT-weighted MEPS-ICAR total family income is very close to the ACS total family income at the mean and closer than the over-the-year estimate at every highlighted percentile. The difference between the MEPS-ICAR estimates targeting an over-the-year employment concept versus a point-in-time concept likely reflect a tendency of workers from lower income families to have shorter job tenures than those from higher income families, with this tendency at least partly being mechanical (i.e., your family income will be lower if you were unemployed for longer in a given year). Even after PIT weighting, the MEPS-ICAR numbers do still tend to be lower than the ACS ones. In addition to issues relating to the IRS Form 1040 match rate, these differences may also reflect underlying differences in how IRS tax filing units correspond with ACS families. Tax filing units can often be smaller than ACS families, especially for low-income families, which would tend to push family income estimates in the MEPS-ICAR downwards. Overall, Table 7's estimates suggest that MEPS-ICAR personal and family income data follow

a distribution similar to that in the ACS, provided one uses the MEPS-ICAR's point-in-time weights to improve the degree of conceptual correspondence between what the MEPS-ICAR and the ACS are measuring.

Commuting Distances for Workers

The final table of worker-level data is Table 8, which compares commute distances calculated for workers in the MEPS-ICAR with those calculated in the 2017 and 2009 National Household Travel Survey (NHTS; U.S. Department of Transportation, 2009, 2017). Means and various percentiles across the commute distance distributions are presented. When considering these numbers, note that the 2009 NHTS top codes its distance to work at a significantly lower threshold than the 2017 NHTS, with this difference accounting for the difference in mean commutes between the two datasets. For the MEPS-ICAR commute numbers, we present estimates both using and not using point-in-time weights. For each, we show two different types of commute distance: one is the commute distance for all workers, while the other is the commute distance for the bottom 90 percent of workers in terms of commute distance. Both measures are included, because the MEPS-ICAR commute data distribution is heavily right-skewed, so viewing the trimmed data can be informative.

Starting with the MEPS-ICAR data, the mean MEPS-ICAR worker lives about 71.82 miles from their job, with the mean being 18.01 miles in the sample trimming the top 10 percent of commutes. After application of point-in-time weights, these means fall to 58.31 and 16.70 miles respectively, perhaps reflecting the fact that some jobs worked for less than a full year may have been worked by individuals who moved in that year. Given that the MEPS-ICAR commute distribution has an extreme right tail of workers with very long commutes, it is important to not just focus on these means, as means are highly sensitive to outliers. The 25th, 50th, and 75th percentile MEPS-ICAR workers live about 2.7, 7.6, and 21.2 miles from their jobs respectively, or about 2.6, 7.2, and 18.9 miles in PIT-weighted terms. The same figures in the trimmed sample are generally similar, though smaller. Compare this to the 2017 and 2009 NHTS surveys, which have their mean workers living 22.32 and 13.35 miles from their workplaces. The two years of data have similar percentile commute distances, with each having workers travel about 4, 9, and 18 miles at the 25th, 50th, and 75th percentiles of the commute distribution respectively. Broadly speaking, the commute distances at percentiles 5, 10, 25, 50, and 75 are quite similar across the MEPS-ICAR estimates (trimmed or untrimmed; PIT weighted or unweighted) and the two NHTS surveys, indicating broad correspondence between the MEPS-ICAR and the NHTS for a large majority of workers. The point-in-times weights do, however, tend to help pull the MEPS-ICAR commute distance figures closer to the NHTS estimates in general. The trimmed MEPS-ICAR distances are also similar to the NHTS commute distances at the mean and the 90th percentile, if not all the way out to the 95th percentile.

There are two key areas of divergence between the NHTS and MEPS-ICAR commute distances. First, in the bottom half of the commute distance, the MEPS-ICAR commute distances tend to be systematically shorter than the NHTS distances. This is likely because worker and establishment locations in the MEPS-ICAR are often only approximate. As a result, the distance between workers and establishments in the same zip code will often be set to 0 in the MEPS-ICAR, pushing down the MEPS-ICAR commute distances by a small amount. Second, even after application of point-in-time weights, the MEPS-ICAR has an extreme right tail of workers with very long commute distances that are not present in the NHTS. This likely results from a few factors. First, the MEPS-ICAR worker residence locations are primarily drawn from Form 1040s. Residence locations for workers that do not match to Form 1040s must be derived from other sources that may be from different data years. Form 1040 residences themselves may be incorrect for workers that leave their MEPS-ICAR jobs and move to a location away from their old job. The MEPS-ICAR may also have difficulties when workers typically commute to work from a residence other than the one listed on their Form 1040 (e.g., a worker might maintain a residence near a natural gas field where they work in North Dakota, while the rest of their family lives in another residence in another state). In these cases, the MEPS-ICAR will estimate very long commute distances, whereas the NHTS will not, because it is a survey-based measure of actual distances traveled to work. These difficulties with unrealistically long commutes in the right tail of the distribution are similar, albeit less severe in some respects, than those found in analyses of commute distance measures derived from similar employer-worker linked datasets, such as Green, Kutzbach, and Vilhuber's (2017) analysis of commute distance data from the LEHD program's Origin-Destination Employment Statistics (LODES) dataset.

Overall, we would characterize these commute distance distributions as being heartening and suggestive of a successful match between MEPS-IC establishments and their workforces, with the trimmed MEPS-ICAR commute distances likely being the most relevant for consideration given certain quality issues with MEPS-ICAR commute figures in the right tail of the distribution. Getting largely appropriate commute distributions is of particular note given that these distances were an input into the match process itself. While the untrimmed MEPS-ICAR commute distances do diverge considerably in the right tail of the distribution from the NHTS commute distances, the reasons this occur do not generally present much cause for broader concern. Moreover, the fact that matched workers can still have very long commute distances is indicative that the matching algorithm's prioritization of matching employment and payroll targets over minimizing commute distances likely struck an appropriate balance.

V. Conclusion

The MEPS-ICAR links survey data on MEPS-IC establishments and their health insurance benefits packages to detailed data on those establishments' workforces, including data on their workers' personal incomes, family incomes, demographic characteristics, and residential locations. The MEPS-ICAR also provides the same information for the workforces of MEPS-IC establishments' parent firms, alongside establishment- and firm-level employee turnover statistics. A key caveat on the MEPS-ICAR data is that while it does provide information about the health insurance benefits choice set that establishments offer to their employees and overall enrollment in each insurance plan, it does not include direct information about which health insurance plan is chosen by particular linked workers. With respect to MEPS-ICAR data quality, match rates between establishments and their workforces are consistently high across nearly all subgroups of establishments, with quality assessment statistics speaking favorably to the reliability of MEPS-ICAR data in terms of employment, payroll, and other characteristics. One important proviso on the quality of the MEPS-ICAR data that analysts should be aware of is that its family income and composition data is derived from IRS Form 1040s. Since the Form 1040s employ definitions of families and certain related concepts (e.g., marital status) that can differ from those in commonly used surveys, analysts should be careful when comparing MEPS-ICAR estimates to estimates from other sources. Analysts should also be sure to use the MEPS-ICAR's point-in-time weights when seeking to directly compare MEPS-ICAR estimates to outside data sources that measure employment conditions at particular points in time rather than over the course of a year.

The MEPS-ICAR presents considerable opportunities for researchers. We highlight a selection of five potential research areas that may particularly benefit from new MEPS-ICAR data:

- Understanding how health insurance offers and benefits vary by worker characteristics would benefit from the MEPS-ICAR's greater demographic detail about establishments' workforces. In particular, where the MEPS-IC was limited to reporting the percentage share of workers aged 50 or over along with the percentage share of workers that are women, the MEPS-ICAR offers information on the full joint distribution of workforce racial/ethnic composition, age, sex, and marital status.
- Research into the compensating differentials associated with employers' health insurance offers should benefit from new data on workers' personal and family incomes.
- Research into how employer-sponsored health insurance offers affect labor mobility (and vice versa) should benefit from the MEPS-ICAR's new measures of employee turnover.

- Research into how employers structure their health insurance benefits packages in response to their workforce's composition should benefit from the MEPS-ICAR's new data on the workforces of MEPS-IC establishments' parent firms. In particular, this new data allows researchers to consider how differences between a firm's overall workforce and its workforce at particular MEPS-IC establishments might affect benefits package offers to, and take-up by, the workers at MEPS-IC establishments.
- Research on how state and national policies affect employers' health insurance offering decisions should benefit from the MEPS-ICAR's new data on workers' residential locations, as well as from other MEPS-ICAR data on workers' family income and characteristics more broadly. This new data should enable researchers to assess which state policies affect a given employer's workforce, to assess workers' Medicaid eligibility, and to assess how a range of other policies (e.g., changes in tax policy, Affordable Care Act subsidy rules) affect employers' workforces.

In addition to creating new opportunities for analysts, the construction of the MEPS-ICAR has also generated a number of benefits for the baseline MEPS-IC survey. First, the MEPS-ICAR's steady-state employment measure derived from tax data when calculating turnover statistics serves as a new, external check on the quality of the data collected by the MEPS-IC survey's employment question. Comparison of the two employment measures points to a generally high degree of correspondence, suggesting the quality of the survey data is high for most establishments. Second, in the process of constructing the MEPS-ICAR, we discovered that the MEPS-IC survey has, in the past, faced difficulty measuring employment for establishments heavily involved in either providing or hiring contract workers. While these problems were not so prevalent as to generate large biases across the full distribution of establishments in the previously mentioned employment data quality check, these issues were responsible for some cases where the two measures diverged significantly. Improvements to the core MEPS-IC employment question have been made to address the discovered issues by clarifying to establishments how to respond to questions with respect to their contract workers and with respect to their workers detailed to worksites that either are not owned by the respondent business or that lack fixed locations.

Overall, the construction of the MEPS-ICAR has yielded dividends for the underlying MEPS-IC survey itself while considerably expanding the range of questions the MEPS-IC survey data can address. Going into the future, we expect the MEPS-ICAR to bear substantial fruit in terms of novel research and further benefits to the underlying MEPS-IC survey.

VI. References

- Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. 2005. "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators." U.S. Census Bureau, LEHD Program Technical Paper No. TP-2006-01.
- Agency for Healthcare Research and Quality (AHRQ). 2005–2017. The Medical Expenditure Panel Survey – Insurance Component. meps.ahrq.gov/survey_comp/ic_technical_notes.shtml
- Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica*, 80: 2369–2429.
- Davis, Karen E. 2018. Sample Design of the 2017 Medical Expenditure Panel Survey Insurance Component. Methodology Report 31, Agency for Healthcare Research and Quality.
- DeSalvo, Bethany, Frank F. Limehouse, and Shawn D. Klimek. 2016. "Documenting the Business Register and Related Economic Business Data." U.S. Census Bureau Center for Economic Studies Working Paper, CES-16-17.
- Graham, Matthew R., Mark J. Kutzbach, and Danielle H. Sandler. 2017. "Developing a Residence Candidate File for Use with Employer-Employee Matched Data." U.S. Census Bureau Center for Economic Studies Working Paper, CES-17-40.
- Green, Andrew S., Mark J. Kutzbach, and Lars Vilhuber. 2017. "Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files." U.S. Census Bureau Center for Economic Studies Working Paper, CES-17-34.
- McCue, Kristin and Martha Stinson. 2019. "Readme_W2." Internal U.S. Census Bureau Center for Economic Studies technical document.
- Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouwiler, and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021.
- U.S. Census Bureau. 2005–2017. The American Community Survey.
- U.S. Department of Transportation, Federal Highway Administration. 2009. National Household Travel Survey.
- U.S. Department of Transportation, Federal Highway Administration. 2017. National Household Travel Survey.
- Wagner, Deborah and Mary Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software." U.S. Census Bureau Center

for Administrative Records Research and Applications Working Paper, CARRA
Working Paper #2014-01.

VII. Tables

Table 1. Establishment-Workforce Successful Match Rates by Assorted Subgroups

	Establishment-Weighted Estimates			Employment-Weighted Estimates		
	% Match	% Match Among Health Insurance Offerors	% Match Among Non-Offerors	% Match	% Match Among Health Insurance Offerors	% Match Among Non-Offerors
All	92.89	94.00	91.73	93.33	93.31	93.46
Years						
2005	94.79***	94.62**	95.02***	94.24**	94.07*	95.40***
2006	95.42***	95.67***	95.11***	94.28**	94.16*	95.05**
2008	91.37***	93.40*	88.74***	93.40	93.85	90.22**
2009	91.53***	93.35*	89.30***	93.34	93.57	91.71***
2010	91.31***	93.16**	89.15***	92.44*	92.51+	92.04***
2011	91.52***	93.20**	89.77***	93.03	93.20	92.07**
2012	91.94***	93.76	90.10***	92.80+	92.86	92.45*
2013	93.24+	94.19	92.29+	93.46	93.37	93.94
2014	92.84	93.99	91.81	93.33	93.42	92.91
2015	93.60***	94.39	92.94***	93.69	93.44	95.01***
2016	93.72***	94.15	93.37***	93.41	93.09	95.07***
2017	93.28	94.11	92.55*	92.54*	92.19**	94.47**
Census Divisions						
New England	94.10***	94.94***	92.96***	95.17***	95.25***	94.56
Middle Atlantic	93.57***	94.47*	92.43*	94.50***	94.50***	94.53**
East North Central	93.67***	94.38*	92.91***	94.60***	94.68***	94.05
West North Central	93.42**	94.41+	92.47**	93.82	93.82	93.86
South Atlantic	92.61+	93.93	91.37	91.74***	91.44***	93.42
East South Central	92.92	93.72	92.03	93.66	93.67	93.61
West South Central	91.84***	92.73***	91.02*	91.99***	91.75***	93.15
Mountain	92.16***	93.38**	91.09*	91.72***	91.43***	93.12
Pacific	92.28***	93.87	90.59***	93.62	93.90*	92.12**
Industry						
Agriculture, Fishing, & Forestry	91.18***	92.88	90.64+	88.81***	87.51**	90.48**
Mining & Manufacturing	93.71**	94.85***	91.69	94.67***	94.70***	94.07
Construction	91.53***	94.05	89.98***	93.88+	94.18*	93.00
Utilities & Transport.	91.87**	92.74**	90.85	94.01	94.14+	92.62
Wholesale	93.47*	94.44+	91.71	92.83	92.72	93.96
Financial Services & Real Estate	92.63	93.06***	91.84	92.90	92.94	92.45+
Retail	93.49***	94.97***	91.67	95.44***	95.65***	94.05*

Professional Services	93.45***	94.20	92.55***	93.29	93.16	94.43***
Other Services	92.60*	93.50**	91.98	91.92***	91.45***	93.28
Asstd. Firm Characteristics						
Single-Estab. Firm	92.94	95.15***	91.70	95.32***	96.27***	93.58*
Multi-Estab. Firm	92.75	92.80***	92.25	92.19***	92.19***	92.47*
For-Profit	92.65***	93.77***	91.54***	92.86***	92.78***	93.32***
Non-Profit	95.62***	96.11***	94.80***	96.18***	96.21***	95.67***
Firm Size						
1 Employee	87.80***	90.13***	87.36***	87.71***	90.13***	87.25***
2-9	93.87***	95.14***	93.23***	94.20***	95.24***	93.55
10-49	95.16***	95.72***	94.14***	95.65***	96.11***	94.68***
50-99	95.02***	95.23***	93.57**	96.02***	96.25***	94.40
100-999	93.68***	93.78	91.69	94.60***	94.72***	91.77
1000+	92.20***	92.20***	92.53	91.51***	91.53***	88.30
Establishment Size						
1 Employee	87.88***	89.64***	87.33***	87.88***	89.64***	87.33***
2-5	93.64***	94.64***	93.03***	93.89***	94.81***	93.27
6-19	94.25***	94.32**	94.11***	94.24***	94.26***	94.20***
20-49	94.26***	94.15	94.79***	94.21***	94.08***	94.86***
50-99	94.39***	94.41*	94.06**	94.49***	94.52***	94.12
100+	94.12***	94.14	93.35	92.41***	92.43***	90.95

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are either representative of MEPS-IC establishments (i.e., establishment-weighted, using the MEPS-IC survey weights estimates) or are representative of employees of MEPS-IC establishments (i.e., employment-weighted, using the MEPS-IC survey weights * MEPS-IC survey reported employment estimates). Match rates shown are successful match rates (i.e., all matches less matches failing to meet minimum quality requirements). The full sample contains 354,000 establishments (328,000 matched and 26,000 unmatched) across 303,000 firms (280,000 matched firms and 23,000 unmatched). Statistical significance indicators are attached to match rate estimates for all sample subgroups. These indicators show results from tests of the hypothesis that the match rate for the subgroup specified by the row is equal to the match rate for all other subgroups combined together. The symbols shown map into p-values as follows: *** for $p < 0.001$, ** for $p < .01$, * for $p < .05$, and + for $p < .1$.

Table 2A. Establishment-Weighted Estimates for Closely Related Employment and Employee Turnover Statistics at the Firm and Establishment Levels

	Establishment-Weighted Estimates								
	Mean	5th	10th	25th	50th	75th	90th	95th	SD
Establishment Employment									
Target Worker Count	29.84	1	1	2	7	21	59	107	186.7
Matched Worker Count	27.37	1	1	3	7	19	53	99	157.6
MEPS-IC Reported Emp.	16.96	1	1	2	4	11	30	56	109.1
Tax-Derived Employment	18.74	1	1	2	5	13	33.5	64	116.8
Estab. Employment Comparison Figures									
Matched Workers / Target Workers	1.054	.6000	.7143	.8400	1.000	1.000	1.500	2.000	.4928
Matched Workers / MEPS-IC Reported Emp.	1.644	.6893	1.000	1.000	1.378	2.000	2.750	3.381	1.057
Tax-Derived Employee Turnover	.4626	.0000	.0000	.0000	.2609	.6286	1.119	1.618	.6676
Firm Employment									
Matched Worker Count	10,040	1	1	3	9	67	7,838	40,660	61,250
MEPS-IC Reported Emp.	7367	1	1	2	6	42	6,500	31,000	43,340
Tax-Derived Employment	6799	1	1	2	6	44	5,054	26,290	42,830
Firm Employment Comparison Figures									
Matched Workers / MEPS-IC Reported Emp.	2.555	.5000	.8587	1.000	1.267	1.808	2.667	3.667	75.53
Tax-Derived Employee Turnover	.4574	.0000	.0000	.0000	.2751	.6207	1.083	1.500	.6459

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are representative of MEPS-IC establishments (i.e., establishment-weighted, using the MEPS-IC survey weights estimates). The sample contains 328,000 establishments across 303,000 firms. The four employment measures shown differ as follows: the number of matched workers is the number of W2s matching to a given establishment (or firm), the target number of worker is the number of W2s the match sought to link to a given establishment, the MEPS-IC reported employment level is the number of workers an establishment (or firm) reports employing during a typical pay period, and the tax-derived employment figure is the number of workers at an establishment (or firm) that remain employed there from one year into the next (i.e., another steady-state employment measure like the MEPS-IC reported total). The target worker counts are derived from a simple machine learning model trained on MEPS-IC data for single-establishment firm slinked to tax records.

Table 2B. Employment-Weighted Estimates for Closely Related Employment and Employee Turnover Statistics at the Firm and Establishment Levels

	Employment-Weighted Estimates								
	Mean	5th	10th	25th	50th	75th	90th	95th	SD
Establishment Employment									
Target Worker Count	1,155	4	10	35	132	539.7	2,130	4,616	5,482
Matched Worker Count	965.3	5	9	31	122	491	1,897	4,086	3,667
MEPS-IC Reported Emp.	718.5	4	6	19	77	337	1,393	3,086	3,127
Tax-Derived Employment	724.4	3.5	7	20.5	80.5	349.5	1,441	3,210	2,630
Estab. Employment Comparison Figures									
Matched Workers / Target Workers	.9330	.6361	.7403	.8085	.9347	1.000	1.148	1.209	.2260
Matched Workers / MEPS-IC Reported Emp.	1.614	.9131	1.000	1.191	1.446	1.848	2.400	3.100	.7151
Tax-Derived Employee Turnover	.5218	.02885	.08410	.1697	.3404	.6575	1.127	1.559	.6189
Firm Employment									
Matched Worker Count	43,060	5	11	55	620	13,560	79,690	223,000	179,100
MEPS-IC Reported Emp.	31,140	4	8	36	514.5	13,850	61,780	150,000	122,900
Tax-Derived Employment	29,430	4	8	37	430.5	9,657	54,130	149,300	125,800
Firm Employment Comparison Figures									
Matched Workers / MEPS-IC Reported Emp.	2.313	.3230	.6133	1.033	1.302	1.711	2.429	3.278	47.30
Tax-Derived Employee Turnover	.5149	.05263	.1064	.1867	.3548	.6502	1.085	1.479	.5638

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are representative of employees of MEPS-IC establishments (i.e., employment-weighted, using the MEPS-IC survey weights * MEPS-IC survey reported employment estimates). The sample contains 328,000 establishments across 303,000 firms. The four employment measures shown differ as follows: the number of matched workers is the number of W2s matching to a given establishment (or firm), the target number of worker is the number of W2s the match sought to link to a given establishment, the MEPS-IC reported employment level is the number of workers an establishment (or firm) reports employing during a typical pay period, and the tax-derived employment figure is the number of workers at an establishment (or firm) that remain employed there from one year into the next (i.e., another steady-state employment measure like the MEPS-IC reported total). The target worker counts are derived from a simple machine learning model trained on MEPS-IC data for single-establishment firm slinked to tax records.

Table 3A. Establishment-Weighted Estimates for Closely Related Payroll Statistics at the Firm and Establishment Levels

	Establishment-Weighted Estimates								
	Mean	5th	10th	25th	50th	75th	90th	95th	SD
Establishment Payroll									
Target Payroll	78,6700	10,930	19,250	45,620	130,200	373,300	1.110 x 10 ⁶	2.345 x 10 ⁶	8.810 x 10 ⁶
Matched Payroll	80,8900	9,901	17,800	45,000	130,100	376,800	1.131 x 10 ⁶	2.397 x 10 ⁶	9.458 x 10 ⁶
Business Register Payroll	81,6300	11,000	19,000	46,000	131,000	375,000	1.118 x 10 ⁶	2.376 x 10 ⁶	1.031 x 10 ⁷
Matched/Target Payroll	1.239	.7827	.8231	1.000	1.000	1.000	1.213	1.377	11.03
Matched/Business Register Payroll	1.122	.7112	.8142	.9858	1.000	1.014	1.195	1.278	155.1
Firm Payroll									
Matched Payroll	3.024 x 10 ⁸	10,420	18,740	49,500	179,400	1.603 x 10 ⁶	1.717 x 10 ⁸	9.412 x 10 ⁸	1.805 x 10 ⁹
Business Register Payroll	3.799 x 10 ⁸	11,000	20,000	50,000	181,000	1.766 x 10 ⁶	2.608 x 10 ⁸	1.401 x 10 ⁹	2.150 x 10 ⁹
Matched/Business Register Payroll	.9773	.5000	.7922	.9778	.9995	1.003	1.029	1.094	1.072

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are representative of MEPS-IC establishments (i.e., establishment-weighted, using the MEPS-IC survey weights estimates). The sample contains 328,000 establishments across 303,000 firms. All payroll estimates shown are in dollars. The target payroll total is the total quantity of payroll the matching algorithm initially sought to match to a given establishment (or firm), while the matched payroll total is the total quantity of payroll (summing across all matched W2s) actually matched to the given establishment (or firm). The Business Register payroll total is the quantity of payroll reported for the establishment (or firm) on the Business Register. The target payroll totals are essentially the same as the Business Register ones, except with some additional data cleaning rules having been imposed.

Table 3B. Employment-Weighted Estimates for Closely Related Payroll Statistics at the Firm and Establishment Levels

	Employment-Weighted Estimates								
	Mean	5th	10th	25th	50th	75th	90th	95th	SD
Establishment Payroll									
Target Payroll	4.304 x 10 ⁷	74,000	150,000	495,600	2.464 x 10 ⁶	1.352 x 10 ⁷	7.325 x 10 ⁷	1.823 x 10 ⁸	2.070 x 10 ⁸
Matched Payroll	4.703 x 10 ⁷	73,040	151,800	508,700	2.571 x 10 ⁶	1.450 x 10 ⁷	7.988 x 10 ⁷	2.022 x 10 ⁸	2.276 x 10 ⁸
Business Register Payroll	4.374 x 10 ⁷	72,000	148,000	491,000	2.484 x 10 ⁶	1.380 x 10 ⁷	7.491 x 10 ⁷	1.863 x 10 ⁸	2.096 x 10 ⁸
Matched/Target Payroll	1.026	.7139	.8007	.8479	1.000	1.000	1.251	1.331	2.903
Matched/Business Register Payroll	11.21	.7292	.7860	.9807	1.001	1.158	1.248	1.431	2649
Firm Payroll									
Matched Payroll	1.095 x 10 ⁹	82,710	198,500	1.142 x 10 ⁶	1.578 x 10 ⁷	4.165 x 10 ⁸	2.423 x 10 ⁹	6.611 x 10 ⁹	3.588 x 10 ⁹
Business Register Payroll	1.349 x 10 ⁹	84,000	202,000	1.205 x 10 ⁶	2.133 x 10 ⁷	6.602 x 10 ⁸	3.364 x 10 ⁹	7.534 x 10 ⁹	4.060 x 10 ⁹
Matched/Business Register Payroll	.9016	.2140	.4315	.8423	.9962	1.001	1.025	1.081	1.698

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are representative of employees of MEPS-IC establishments (i.e., employment-weighted, using the MEPS-IC survey weights * MEPS-IC survey reported employment estimates). The sample contains 328,000 establishments across 303,000 firms. All payroll estimates shown are in dollars. The target payroll total is the total quantity of payroll the matching algorithm initially sought to match to a given establishment (or firm), while the matched payroll total is the total quantity of payroll (summing across all matched W2s) actually matched to the given establishment (or firm). The Business Register payroll total is the quantity of payroll reported for the establishment (or firm) on the Business Register. The target payroll totals are essentially the same as the Business Register ones, except with some additional data cleaning rules having been imposed.

Table 4A. Closely Related Employment, Employee Turnover, and Payroll Statistics at the Firm and Establishment Levels by Single-Establishment Firms

	Single-Establishment Firms							
	Establishment-Weighted Estimates				Employment-Weighted Estimates			
	Mean	25th	50th	75th	Mean	25th	50th	75th
Establishment Employment								
Target Worker Count	15.04	2	4	13	194.1	12	38	118
Matched Worker Count	13.84	2	5	12	172.5	11	33	106
MEPS-IC Reported Emp.	8.703	2	3	7	105.8	7	22	68
Tax-Derived Emp.	9.331	2	4	8.5	109.1	8	22.5	69
Establishment Emp. Comparison Figures								
Matched Workers / Target Workers	1.072	.8571	1.000	1.010	.9391	.8000	.9019	1.000
Matched Workers / MEPS-IC Reported Emp.	1.575	1.000	1.250	1.833	1.591	1.148	1.410	1.833
Tax-Derived Employee Turnover	.4207	.0000	.2222	.5556	.5477	.1783	.3750	.7022
Establishment Payroll								
Target Payroll	353,400	34,880	92,810	258,000	4.505 x 10 ⁶	192,800	681,200	2.558 x 10 ⁶
Matched Payroll	346,100	34,500	91,450	254,100	4.470 x 10 ⁶	188,800	667,100	2.492 x 10 ⁶
Business Register Payroll	354,800	35,000	93,000	258,000	4.497 x 10 ⁶	193,000	681,200	2.548 x 10 ⁶
Matched/Target Payroll	1.159	1.000	1.000	1.000	1.067	1.000	1.000	1.000
Matched/Business Register Payroll	1.007	.9917	.9999	1.004	1.012	.9970	.9999	1.001

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are either representative of MEPS-IC establishments (i.e., establishment-weighted, using the MEPS-IC survey weights estimates) or are representative of employees of MEPS-IC establishments (i.e., employment-weighted, using the MEPS-IC survey weights * MEPS-IC survey reported employment estimates). The sample contains 185,000 establishments that are part of single-establishment firms and 143,000 establishments that are part of multi-establishment firms; only estimates for single-establishment firms are shown here. All payroll estimates shown are in dollars. For more on the differences between the different employment and payroll figures, please see the notes to Tables 2A, 2B, 3A, and 3B.

Table 4B. Closely Related Employment, Employee Turnover, and Payroll Statistics at the Firm and Establishment Levels for Multi-Establishment Firms

	Multi-Establishment Firms							
	Establishment-Weighted Estimates				Employment-Weighted Estimates			
	Mean	25th	50th	75th	Mean	25th	50th	75th
Establishment Employment								
Target Worker Count	68.80	6	20	57	1,723	90	268	1,040
Matched Worker Count	62.99	6	19	51	1,435	82	250	935.1
MEPS-IC Reported Emp.	38.70	4	10	25	1,081	47	170	678.6
Tax-Derived Emp.	43.51	4.5	12	32	1,089	53.5	174.4	699.5
Establishment Emp. Comparison Figures								
Matched Workers / Target Workers	1.004	.8158	1.000	1.000	.9293	.8163	.9565	1.000
Matched Workers / MEPS-IC Reported Emp.	1.826	1.300	1.643	2.000	1.628	1.211	1.466	1.853
Tax-Derived Employee Turnover	.5728	.1667	.4000	.7729	.5065	.1672	.3241	.6255
Establishment Payroll								
Target Payroll	1.927 x 10 ⁶	120,900	288,700	847,000	6.586 x 10 ⁷	1.197 x 10 ⁶	5.646 x 10 ⁶	3.173 x 10 ⁷
Matched Payroll	2.027 x 10 ⁶	124,900	301,900	908,000	7.223 x 10 ⁷	1.305 x 10 ⁶	6.184 x 10 ⁶	3.512 x 10 ⁷
Business Register Payroll	2.031 x 10 ⁶	124,000	292,900	856,800	6.697 x 10 ⁷	1.187 x 10 ⁶	5.752 x 10 ⁶	3.242 x 10 ⁷
Matched/Target Payroll	1.448	.8121	.9415	1.142	1.002	.8060	.9177	1.041
Matched/Business Register Payroll	1.425	.8543	1.042	1.206	17.25	.9378	1.069	1.209

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: Estimates are either representative of MEPS-IC establishments (i.e., establishment-weighted, using the MEPS-IC survey weights estimates) or are representative of employees of MEPS-IC establishments (i.e., employment-weighted, using the MEPS-IC survey weights * MEPS-IC survey reported employment estimates). The sample contains 185,000 establishments that are part of single-establishment firms and 143,000 establishments that are part of multi-establishment firms; only estimates for multi-establishment firms are shown here. All payroll estimates shown are in dollars. For more on the differences between the different employment and payroll figures, please see the notes to Tables 2A, 2B, 3A, and 3B.

Table 5. Supplementary Match Quality Assessment Regressions

	MEPS Emp. on Matched Workers	Target Workers on Matched Workers	Business Register Payroll on Matched Payroll	Target Payroll on Matched Payroll	MEPS % Women on Matched % Women	MEPS % Age 50+ on Matched % Age 50+	MEPS Emp. on Matched Workers	Business Register Payroll on Matched Payroll
Establishment-Weighted Estimates								
Coefficient	.6404***	1.133***	.8902***	.8698***	.7692***	.6597***	.6775***	1.150***
Standard Error	(.01314)	(.01660)	(.01673)	(.01612)	(.002045)	(.002997)	(.003285)	(.004919)
R ²	.8595	.9170	.6685	.8731	.8489	.7140	.9190	.9341
Employment-Weighted Estimates								
Coefficient	.7663***	1.369***	.8608***	.8490***	.8128***	.6085***	.6733***	1.080***
Standard Error	(.08270)	(.1344)	(.02899)	(.02803)	(.001848)	(.002615)	(.003410)	(.003814)
R ²	.8171	.8459	.8793	.8769	.9174	.7953	.9650	.9195

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records, 2005–2017 excluding 2007.

Notes: All estimates shown are from regressions of the first variable listed in the column title (either a target quantity, an estimate from the MEPS-IC survey, or an estimate from the Business Register) on the second variable listed in the column title (a measure from the matched sample of workers). Some regressions use establishment-level variables while others use firm-level variables. The top panel of the table uses only MEPS-IC survey weights to obtain establishment-weighted estimates while the bottom panel uses those same survey weights multiplied by MEPS-IC survey reported employment to obtain employment-weighted estimates. Standard errors associated with each coefficient are shown under the coefficients in parentheses. Coefficients are marked with statistical significance indicators which represent the following: *** for $p < 0.001$, ** for $p < .01$, * for $p < .05$, and + for $p < .1$. The sample contains 328,000 establishments across 303,000 firms. All payroll estimates shown are in dollars. For more on the differences between the different employment and payroll figures, please see the notes to Tables 2A, 2B, 3A, and 3B.

Table 6. MEPS-ICAR Demographic, Marital, and Family Characteristics vs. American Community Survey Benchmarks

	MEPS-ICAR Mean	MEPS-ICAR PIT-weighted Mean	ACS Mean
Match Statistics			
Form 1040 Match Failure Rate	.08392	.0628	---
Decennial Census Match Failure Rate	.06358	.05173	---
Demographic Characteristics (Decennial Census Derived)			
Age	37.87	40.27	41.13
Share Women	.4946	.4929	.4734
Age (Women Only)	37.66	40.08	40.99
Age (Men Only)	38.07	40.46	41.26
Hispanic	.1460	.1359	.1568
Non-Hispanic White	.6740	.6962	.6568
Non-Hispanic Black	.1328	.1167	.1111
Non-Hispanic Asian	.04838	.05189	.0531
Non-Hispanic Other	.008942	.007920	.0222
Hispanic Female	.06883	.06456	.0668
Non-Hispanic White Female	.3296	.3380	.3105
Non-Hispanic Black Female	.07258	.06491	.0597
Non-Hispanic Asian Female	.02445	.02611	.0253
Non-Hispanic Other Female	.004483	.003987	.0111
Hispanic Male	.07716	.07138	.0900
Non-Hispanic White Male	.3444	.3583	.3463
Non-Hispanic Black Male	.06017	.05182	.0514
Non-Hispanic Asian Male	.02393	.02579	.0277
Non-Hispanic Other Male	.004459	.003933	.0111
Marital Status & Family Composition (Form 1040 Derived)			
Single Filing Status/Single without Kids (ACS)	.4035	.3643	.3564
Married Filing Status/Married (ACS)	.4433	.4944	.5435
Widow with Dependents	.0004227	.0004345	---
Household Head Filing Status/Single with Kids (ACS)	.1528	.1409	.1002
Child at Home Exemptions Claimed/Number of Children in Household (ACS)	.7151	.7346	.7898
Child Away from Home Exemptions Claimed	.004801	.005036	---

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records (MEPS-ICAR) and American Community Survey (ACS), 2005–2017 excluding 2007.

Notes: MEPS-ICAR estimates are calculated at the matched-worker level using MEPS-IC survey weights from an overall sample of 56,030,000 observations, containing one observation per worker observed at a MEPS-IC employer over the course of the entire year. MEPS-ICAR Point-in-Time (PIT) weighted estimates are similar, but they are estimated using a set of weights that target employment

at MEPS-IC employers at an average point-in-time, doing so by weighting each observation by the share of the year that the given individual spent working for their MEPS-IC employer. Estimates for Decennial Census and Form 1040 derived variables are for only the subset of workers successfully linked to those data sources. American Community Survey data is drawn from the collection of 1 percent samples for all of the listed years, limiting to just persons in the labor force that have had a job at some point in their lives, do not live in group quarters, do not work in the public sector, are not in the armed forces, and do not report having been continuously unemployed for 5 or more years.

Table 7. MEPS-ICAR Age and Income Levels vs. American Community Survey Benchmarks

	Mean	5th	10th	25th	50th	75th	90th	95th
Age								
MEPS-ICAR	37.87	18	20	25	36	49	59	63
MEPS-ICAR (PIT-Weighted)	40.27	19	21	28	40	52	60	64
ACS	41.13	20	22	29	41	52	60	64
Personal Wage Income								
W2 Pay for MEPS-IC Job	29,550	288	731	3,152	14,020	36,680	68,270	97,820
W2 Pay for MEPS-IC Job (PIT)	40,350	1,311	2,800	9,277	25,300	48,360	83,440	117,600
ACS Wage and Salary Income	38,390	0	1,000	11,000	28,000	50,000	85,000	115,000
Family Total Money Income								
1040 Total Money Income	66,790	4,424	7,731	17,170	37,310	76,040	131,600	188,600
1040 Total Money Income (PIT)	77,710	6,781	11,290	23,040	46,030	86,750	146,000	210,300
ACS Total Family Income	79,526	10,300	18,000	35,000	64,800	107,000	165,000	220,000

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records (MEPS-ICAR) and American Community Survey (ACS), 2005–2017 excluding 2007.

Notes: MEPS-ICAR estimates are calculated at the matched-worker level using MEPS-IC survey weights from an overall sample of 56,030,000 observations, containing one observation per worker observed at a MEPS-IC employer over the course of the entire year. MEPS-ICAR Point-in-Time (PIT) weighted estimates are similar, but they are estimated using a set of weights that target employment at MEPS-IC employers at an average point-in-time, doing so by weighting each observation by the share of the year that the given individual spent working for their MEPS-IC employer. Estimates for Decennial Census and Form 1040 derived variables are for only the subset of workers successfully linked to those data sources. American Community Survey data is drawn from the collection of 1 percent samples for all of the listed years, limiting to just persons in the labor force that have had a job at some point in their lives, do not live in group quarters, do not work in the public sector, are not in the armed forces, and do not report having been continuously unemployed for 5 or more years. All income numbers are in dollars.

Table 8. MEPS-ICAR Commute Data vs. National Household Travel Survey Benchmarks

	Mean	5th	10th	25th	50th	75th	90th	95th
MEPS-ICAR Commute Distances								
Commute Distance	71.82	.06937	.5296	2.713	7.564	21.80	152.7	263.1
Commute Distance (Top 10% Trimmed)	18.01	.05571	.4128	2.418	6.501	15.72	44.02	99.29
MEPS-ICAR Commute Distances (PIT-Weighted)								
Commute Distance	58.31	.05263	.4350	2.605	7.173	18.86	113.7	197.6
Commute Distance (Top 10% Trimmed)	16.70	.04301	.3365	2.366	6.372	14.88	37.42	87.89
National Household Travel Survey Commute Distances								
Distance to Work (2017 NHTS)	22.32	0.86	1.64	3.94	9.18	18.12	30.96	44.15
Distance to Work (2009 NHTS)	13.35	0.78	1.67	4.00	9.00	18.00	30.00	38.00

Source: Medical Expenditure Panel Survey – Insurance Component with Administrative Records (MEPS-ICAR), 2005–2017 excluding 2007, and National Household Travel Survey, 2009 and 2017.

Notes: MEPS-ICAR estimates are calculated at the matched-worker level using MEPS-IC survey weights from an overall sample of 56,030,000 observations, containing one observation per worker observed at a MEPS-IC employer over the course of the entire year. MEPS-ICAR Point-in-Time (PIT) weighted estimates are similar, but they are estimated using a set of weights that target employment at MEPS-IC employers at an average point-in-time, doing so by weighting each observation by the share of the year that the given individual spent working for their MEPS-IC employer. National Household Travel Survey is based on all workers in the United States. All distances shown are in miles.