Prospects for Combining Survey and Non-Survey Data Sources to Improve Estimated Counts of Certain Work-Related Injuries November 2017

Brooks Pierce
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC 20212

Abstract
The Survey of Occupational Injuries and Illnesses (SOII) is the primary survey of nonfatal work-related injuries and illnesses in the U.S. Recently the Occupational Safety and Health Administration (OSHA) has instituted new rules governing employer report of work-related injuries, raising the possibility that OSHA administrative data could be combined with SOII survey data to improve estimates of injury risks in the workplace. This work describes the possibilities and practical challenges of some potential approaches to such a data combination.

Key words: data combination, administrative data, work-related injuries

1. Introduction

Under the Occupational Safety and Health Act of 1970, many employers are required to record work-related injuries and illnesses that occur throughout the year. Under the Act, the Occupational Safety and Health Administration (OSHA) was created to enforce rules and regulations governing workplace safety, and provision was made for collection of injury and illness data recorded by employers. To carry out this task the BLS conducts the annual Survey of Occupational Injuries and Illnesses (SOII). The microdata the SOII collects are essentially OSHA-required forms that log occurring injuries and illnesses.

In 2016, the Occupational Safety and Health Administration (OSHA) issued a new rule mandating certain employers to electronically submit injury and illness data directly to OSHA. Employers in non-exempt industries had previously only been required to keep a log of their workplace injuries and illnesses on forms that the employer maintained. In principle the new OSHA electronic reporting requirements could provide complete administrative records – precisely the forms SOII seeks – for a large part of the SOII scope.

Although this regulatory change could provide an informational windfall to the SOII, it also would present certain challenges. Complete reporting to OSHA is unlikely since employers may be unaware of the new reporting rules. Reporting propensities may vary with observable and unobservable factors, and would likely change through time. Data obtained through OSHA would presumably be subject to a variety of recording errors, some of which would differ from those typically encountered by the SOII. Finally, in order to elicit complete information, respondents to BLS surveys are assured confidentiality under CIPSEA. OSHA is a regulatory agency and plans to post the electronic records publicly. The lack of confidentiality may have an impact on OSHA record quality, for example by inducing under-reporting of injury risk to OSHA. This and other mode effects could lead to bias in the OSHA-collected records and complicate attempts at combining SOII survey data with OSHA-collected records.

Responsible use of OSHA-provided administrative records therefore requires some mechanism to combine records from the different sources and to quantify potential biases in the newly acquired administrative records. This paper describes some relatively straightforward approaches to data combination in this particular context, and some issues that may arise in combining records collected through SOII and OSHA.

Section 2 describes relevant details of OSHA reporting rules and forms; and SOII sampling, data collection, and estimation. Section 3 describes simulations of various data combination attempts under

different reporting selection mechanisms. The data combination approaches are simplistic and transparent adaptations from literatures on dual frame estimation (Lohr and Brick 2012). They are intended to be implementable within current production systems, and so do not require record linkage. Section 4 describes some practical issues that may arise. Section 5 describes record linkage experiments using other OSHA reports and BLS data. The experiments attempt to gauge how successful record linkage between OSHA electronic records and BLS data might be.

2. Background on OSHA reporting rules and SOII processes

2.1 OSHA reporting requirements

OSHA rules currently require certain private sector employers to maintain records of workplace injuries and illnesses, and to post annual summaries for employee viewing. Records are to be retained for a period of time, and are subject to request and inspection by OSHA. Separate records are to be maintained for separate establishments within a firm. Companies with 10 or fewer employees at all times during a calendar year are exempt from record-keeping requirements, as are business establishments within specified industries (generally those with lower workplace injury risk). Regulations and subsequent interpretative findings specify general recording criteria, employee coverage, determination of work-relatedness, and other factors.[1]

Employers must maintain three sets of forms. OSHA form 300 is a log listing basic information on each recordable case. Information includes the employee's name and job title, the date of the incident, a brief description of the incident, and information on the type of injury or illness (for example, whether the injury resulted in time away from work). Form 300A is a summary of the individual case records from form 300. This provides annual totals by injury or illness type along with establishment location and industry, and the establishment's average annual employment and total hours worked by all employees. Form 301 provides further details on the employee and incident for each recordable case.[2,3]

The new OSHA rules require employers to electronically submit these forms. The requirement applies to business establishments with 250 or more employees in those industries required to maintain forms. For a subset of high injury rate industries, business establishments with 20-249 employees must submit the Summary form 300A, but not forms relating case level detail. As of this writing, it is unclear if the new OSHA rules will take effect as originally constituted; they are under litigation and subject to revision by OSHA. However, OSHA has constructed a portal for receiving submitted information and at least by that measure the implementation does not hinge on technical considerations.

2.2 SOII

The SOII is designed to collect occupational injury and illness data directly from employers using a random sample of over 250,000 establishments in private industry and state and local governments. The SOII sample is stratified by state, establishment size class as determined by establishment employment, detailed industry, and ownership group (private, state government, local government). For efficiency reasons the SOII disproportionately samples from strata covering larger establishments and higher injury rate industries (Selby et al. 2008). As these subpopulations are subject to the new OSHA rules, some form

[1] See https://www.osha.gov/recordkeeping/.
[2] There are additional OSHA requirements for reporting certain especially severe injuries; the record linkage attempts at the end of this paper use data generated from that requirement.
[3] Some states administer their own work injury programs, under federal OSHA oversight. Establishments in these state-plan states must satisfy equally rigorous recordkeeping rules but would not electronically submit to federal OSHA.

of data combination could substantially reduce SOII respondent burden, for example through reduced SOII sample size.

Just prior to the start of the year, BLS sends notification and recordkeeping information to establishments selected to be in the sample for the upcoming year. Survey response is required by law, even for establishments otherwise exempt from OSHA recordkeeping requirements. Over the course of the survey year, selected employers record injuries and illnesses in their OSHA logs. For certain injuries and illnesses, those resulting in days away from work (DAFW), employers also record case-level details such as the circumstances surrounding the occurrence and demographic information on the affected employee.[4] In the following January, BLS requests that the selected establishments report this information to BLS. Data collection continues in the first half of the year and estimates are published in the fall.

The SOII publishes a very large number of direct estimates on narrowly defined domains. One set of estimates reports statistics derived from establishment-level totals for hours worked and injuries and illnesses of various types. These statistics are injury and illness totals and mean incidence rates produced along establishment-level dimensions such as state, industry, and establishment size class; distributional information for incidence rates is also tabulated. A second set of estimates reports statistics derived from case-level details about the affected workers, their jobs, and the circumstances surrounding incidents. SOII case-level estimates are primarily restricted to cases which result in days away from work.[5]

2.3 Extent of overlap between SOII and OSHA records

Table 1 gives statistics on the prospective impact in SOII of the electronic reporting rule. Statistics are the percentages of establishments, employment, and case totals in the 2015 SOII private sector sample subject to the additional OSHA reporting requirements (had they been in force at the time). Establishments are subject to either no additional reporting, reporting of summary totals only, or reporting of both summary totals as well as case level information. Distributions are given for weighted as well as unweighted statistics. Unweighted statistics are useful in gauging the resource impact to SOII; weighted statistics are more relevant for gauging the breadth of the rules changes in the economy.

Approximately 40 percent of the establishment records in the SOII would be subject to some form of electronic reporting. One could possibly view this as an upper bound estimate for the cost savings realized by SOII from fully utilizing an influx of OSHA administrative data. The respondent burden is concentrated among larger business establishments in high injury rate sectors, and the weighted statistics indicate the burden falls on less than 1 percent of establishments for reporting case level information and less than 8 percent of establishments for reporting summary establishment level information.

In terms of the workforce affected, the weighted statistics indicate that about 42 percent of the private sector employees within the SOII scope work in establishments subject to the regulatory changes. These 42 percent account for 65 percent of all recordable injury and illness cases. That is, nearly two-thirds of the primary, top-line private sector statistic produced by SOII is attributable to business units subject to the OSHA electronic reporting requirements.

Because the SOII produces a very large number of estimates using reports on cases with days away from work, it is useful to separately look at coverage rates at the case level. Here the relevant group is the subset of business establishments reporting case records. These are the largest business units (250 or more employees) in non-exempt industries. Looking at the weighted statistics, about 25 percent of cases with

[4] Sampled establishments generally report the census of DAFW cases, but for establishments with many DAFW cases the SOII subsamples at the case level. The SOII also collects information on cases that result in days of job transfer and restriction (DJTR), but currently only for a subset of industries.
[5] OSHA case level records would include these cases as well as cases that result in job transfer or restriction.

days away from work (DAFW) and about 35 percent of cases with job transfer or restriction (DJTR) occur in these establishments.[6] This implies, for example, that any mode effects in reporting DAFW case characteristics would potentially affect 25 percent of the data underlying SOII case estimates.

The unweighted statistics for DAFW and DJTR cases indicate that, roughly speaking, 60-70 percent of case records would potentially come from this subset of establishments. This indicates a substantial respondent burden for some establishments. The SOII reduces respondent burden for case-level reporting by subsampling cases. Furthermore, the SOII currently collects DJTR cases only for a subset of industries. Therefore the 62.7 percent (DAFW) and 69.6 percent (DJTR) statistics in panel A do not measure the fraction of current SOII case level records potentially supplanted by OSHA administrative records. These unweighted statistics are however relevant to potential SOII costs should SOII process all such administrative records. As an example, SOII codes occupational classifications from reported job titles for each case record. Part of this coding relies on automated procedures but a substantial part relies on human coding. SOII would presumably carry out similar coding for all OSHA-reported case records it uses. Therefore any source-specific effects – say, a delayed arrival of OSHA administrative records to SOII – would potentially have a large impact on SOII costs.

3. Simulations incorporating OSHA records

We do not yet have OSHA administrative records generated via the new electronic record-keeping rules. In lieu of actual data, I show some simulations of outcomes under various scenarios. I focus on strategies that seek to utilize all available OSHA data records without having to identify SOII respondents in the OSHA source.

The simulations are relatively straightforward. First I constructed a universe frame, which includes both establishment level and case level information. Stratified random samples of establishments on the frame stand in for SOII samples. A known subset of establishments on the frame is subject to OSHA electronic reporting, but only some establishments required to submit do so; the probability an establishment reports is assumed to follow various alternative rules. Finally, I combine the SOII and OSHA records to produce a given set of estimates, using various different methods. Empirical mean squared error statistics for each estimate follow from repeated application of the SOII sampling and OSHA reporting selection mechanisms. MSE statistics can then be compared across estimation methods for each postulated OSHA reporting selection mechanism.

3.1 Frame generation

I pooled all private sector establishment-level responses from the 2013-2015 SOII samples. For SOII strata where the pooled sample sizes exceeded the actual frame unit size, I discarded responses at random to achieve frame size.[7] Then, within the remaining strata where the pooled sample sizes fell short of actual frame size, I duplicated the responses multiple times until achieving a size consistent with the actual frame. Because SOII samples at much higher rates for large establishments and high-injury sectors, this duplicating of responses is relevant primarily for the non-OSHA portion of the frame. I take the resulting set of establishment level information to be my frame for the SOII; this information includes all

[6] The greater fraction for DJTR cases suggests that this subset of business establishments more actively pursues policies of re-tasking injured workers to other duties. The large percentages for DJTR cases also suggests that the SOII might be induced to produce a broader array of estimates for such cases, were the SOII to intake OSHA's electronically reported DJTR case records. This could have further implications, such as an incentive to substitute toward additional DJTR collection within the SOII proper.

[7] The SOII frame contains many establishments that would be nonviable units were they sampled – for example, many units can be expected to be out of business. I take frame size to be the number of viable units within the SOII strata as estimated using 2015 SOII data.

fields from the OSHA summary form 300A. To obtain case-level information consistent with this constructed frame I retrieved the SOII DAFW case reports relevant to each establishment-year pairing on the frame.[8]

3.2 Simulated SOII sampling and post-sampling processing

For the purposes of these simulations, sampling strata are cells determined by state, industry as defined by 3-digit NAICS codes, and establishment employment size groups. These defined strata are coarser than actual SOII strata, particularly along the industry dimension. I used the pooled 2013-2015 SOII samples to derive average annual sampling rates for each defined strata. Each simulation applied these sampling rates to the constructed frame; on average this results in SOII simulated samples of approximately 180,000 establishment units and approximately 200,000 DAFW case records. All sampled units are assumed to respond fully. Therefore post-sampling processing consists only of setting an establishment-level sampling weight equal to the strata's frame size divided by its sample size. The establishment-level weight is carried over to each case record.[9]

3.3 Selection mechanisms into administrative reporting

I assume that larger establishments in higher injury rate industries are directed to submit records to an administrative authority. I set the reporting size cutoff at 50 employees and further restrict reporting to those three-digit industries generally subject to the actual new electronic reporting rules. Establishments not required to submit records do not do so. Some establishments required to report also do not do so; various models for the probability of reporting are possible. If an establishment reports it does so accurately and fully, including all DAFW case record details.[10]

There are three different models for the probability of reporting. One sets the probability of report from each stratum equal to a randomly drawn constant. That is, the probability of report is determined by frame characteristics and is the same for all units in a given stratum. The second sets the probability of reporting to be an increasing function of the number of total cases in the establishment. The third sets the probability of reporting to be a function of the case profile among DAFW cases in the establishment: probabilities of reporting are higher for establishments with a worker experiencing an amputation. The parameters of these three selection mechanisms do not result in similar amounts or patterns of under-reporting to the administrative authority. They are designed only to highlight the challenges faced by different estimation approaches in combining the data. I will refer to these selection mechanisms as "selection on characteristics", "selection on total cases", and "selection on case profiles", respectively. Details on the probabilities are listed in an attachment.

3.4 Simulation and data combination

Each simulation draws a SOII sample from the frame, and also generates a set of establishments reporting to the administrative authority, as described above. For domains not subject to administrative reporting, estimates rely solely on the SOII sample data. For domains subject to administrative reporting, estimates must combine the information from the two sources in some manner. A traditional dual frame approach

[8] Due to partial nonresponse and case subsampling, a SOII respondent may have fewer DAFW case records than its reported DAFW case total. The SOII constructs a case-level adjustment factor to account for such instances. In my pretend frame I maintain that adjustment factor rather than repeat the individual case records to reach the establishment's reported total. Case records in the simulated samples also maintain this adjustment factor.

[9] As per the previous footnote, the case level weights equal the establishment level weight times any relevant case subsampling adjustment factor.

[10] These parameters differ from the actual OSHA electronic reporting rules mainly in that they eliminate the distinction between establishment summary-only reporters and case records reporters, while setting the establishment size cutoff between that relevant for summary reporting (in some instances 20+ employees) and case record reporting (250+ employees).

takes the estimator to be a convex combination of the source-specific estimators. Letting domain be indexed by *d*, establishments in SOII be indexed by *i*, and establishments reporting to OSHA be indexed by *j*, one might form

$$\hat{Y}_d = \theta_d \sum_{i \in SOII} I_{id} w_i y_i + (1 - \theta_d) \sum_{j \in OSHA} I_{jd} w_j y_j \quad (1)$$

where the I_d's are indicator functions for domain membership, the *w*'s are appropriate weights, and the *y*'s are outcomes such as the establishment's number of total cases. The separate indices *i* and *j* are meant to stress that a given establishment need not be linked across sources to form estimates; the θ and (1-θ) weights prevent double-counting. Case-level estimates are formed analogously, but with an additional indicator for whether or not the case has the given characteristic (for example, an indicator for an amputation case). This approach is exceedingly convenient in that with proper weights modifications one can simply append microdata from the different sources and operate on the pooled collection of records.

In the simulations I calculated estimates for total hours worked, total recordable cases, DAFW cases, DJTR cases, and also for two types of particularly severe DAFW cases, those that result in an overnight hospitalization and those that result in an amputation. The domains are three-digit NAICS private sector industry groups, plus an aggregate total, for the US. MSE statistics are calculated based on 500 simulations.

There are several different approaches to operationalizing (1). One approach is to naively insert the OSHA data without weights, analogous to setting the $w_j = 1$.[11] Different variants of this naïve approach correspond to different values for the θ_d: ignoring OSHA data; ignoring SOII data; equally weighting the two sources; and, taking θ_d to be the fraction of establishment units in domain *d* coming from the SOII data. These are not reasonable strategies under the assumptions but they usefully document baselines, and are potentially useful strategies under a mature regime with more complete reporting. The SOII in fact relies on external administrative records for railroad and mining sector data, treating these external records as complete censuses for their respective populations, so these strategies have precedence in the SOII.

A second approach is to set weights w_j via stratifying to cell establishment counts. This stratification will eliminate biases associated with under-reporting where the reporting depends entirely on the observable characteristics determining the strata cells. For this approach (and the others described below) I take θ_d to be the fraction of establishment units in domain *d* coming from the SOII data.

A third approach is to make further adjustments based on reported outcome variables. This approach starts with stratified OSHA weights as above, and further adjusts them based on reported outcomes in the two sources. I look at two variants of this approach. One is a final post-stratification or benchmarking of the OSHA weights so that weighted establishment counts within cells defined by injury rates are equal in the two sources. This process renders OSHA data less helpful for national level estimates, but (possibly) more helpful for smaller domain estimates or estimates of injury subsets. I will refer to this as "calibration". The second variant is to adjust the stratified OSHA weights with a regression coefficient from a regression model of domain-level estimates from the two sources. That is, if the regression equation is

$$\hat{Y}_d^{SOII} = \alpha + \beta \hat{Y}_d^{OSHA} + \epsilon_d \quad (2)$$

[11] In all calculations the *w* weights reflect case subsampling adjustments described earlier, and for SOII the weights also reflect stratification to frame establishment counts.

then the OSHA stratified weights are adjusted by the estimate for β. I refer to this as a Fay-Herriot approach (Fay and Herriot 1979, Ybarra and Lohr 2008).

Worthy alternatives which I have not yet investigated include calibration to multiple outcomes (Sarndal 2007) and a reweighting approach using propensity scores (Elliot and Davis 2005, reminiscent of Dinardo, Fortin and Lemieux 1996).

3.5 Simulation results

In general, the naïve approaches without weights adjustments in the OSHA administrative records perform poorly. This is unsurprising given their large uncorrected biases. The lowest MSE estimator among the naïve approaches involves setting θ_d in equation (1) to be the fraction of establishment units in domain *d* coming from the SOII data.

Accordingly, stratifying the OSHA source weights to frame control totals is hugely beneficial. For simulations where the probability of reporting to OSHA depends entirely on cell membership, stratifying the weights in this way eliminates bias and reduces MSEs below SOII-only levels. For simulations where the probability of reporting depends on outcome variables (here, total cases), stratification is still quite helpful but does not entirely eliminate bias. In these situations either the calibration or Fay-Herriot regression adjustment methods typically further improve MSEs; the data showed a slight preference for the calibration method.

Tables 2 and 3 give examples for a few manufacturing industries. Table 2 gives results where the selection mechanism is on observable characteristics only; table 3 gives results where the selection mechanism is on the outcome variable, total cases. The tables give MSE statistics, specifically, root MSE measured as a percentage of the true population value, for several outcome variables. The outcome variables include important SOII fields such as DAFW, DJTR, and total case counts; and, full-time equivalent employment (an input for SOII incidence rate estimates). I also include statistics for two less frequent but very severe outcomes, DAFW cases that result in overnight hospitalization, or in amputation. The prevalence in actual data of overnight hospitalizations among private sector DAFW cases is roughly 1 in 20. The analogous prevalence of amputations is roughly 1 in 200.

These industries had above-average incidence rates in 2015, based on actual published numbers. Of the industries listed, employment is highest in food manufacturing, next highest in beverage and tobacco manufacturing, and lower in the two textile-related industries. The relative sizes of the industries are reflected in the SOII-only MSE statistics in these tables.

The patterns in table 2 suggest that combining the administrative and survey data is not a good strategy without adjusting the administrative source weights to control totals. In these experiments the under-reporting is severe enough that discarding the administrative records is generally preferable to using them in naïve combination. The exceptions are for the severe but infrequent DAFW incidents in small industries, where variance issues dominate. Some of the amputations estimates in table 2 are unpublishable according to SOII reliability criteria with SOII data alone, but publishable with the addition of the administrative data.

Figure 1 shows MSE statistics for the total cases column for all 49 3-digit industries where there are two data sources in the simulations. This is just a visual summary of the first column in table 2, but for all industries.

Table 3 gives MSE statistics under an alternative reporting selection mechanism, where the probability of reporting to OSHA increases with the establishment's total cases. Because establishment characteristics such as industry and establishment size correlate with case totals, stratified weights will partially correct the resulting bias. However, the bias correction is incomplete and for many estimates the MSEs are larger when using the administrative data with stratified weights, than when discarding the administrative data

altogether. The further weights adjustments either through the Fay-Herriot adjustment or the calibration approach generally result in lower MSE estimates. In particular the SOII-only estimates are generally not quite competitive with the calibration approach, at least for these example industries. Figure 2 summarizes the total cases column, showing all industries. There appears to be a negligible benefit to the calibration approach, as compared to ignoring the OSHA data entirely. Stratification does not correct the OSHA bias sufficiently.

Table 4 gives MSE statistics for the severe DAFW cases, under a third reporting selection mechanism. Here the experiments presume that the reporting probability depends on the case profile in the establishment. In particular, an establishment experiencing an amputation is assigned a larger than average probability of reporting. Therefore any reweighting of the administrative data based on frame characteristics or reported total cases will overinflate amputations cases and very slightly underinflate other outcomes.[12] The bias should generally be at least partly offset by variance considerations. Table 4 shows the entire private sector aggregate. Figures 3 and 4 give distributions across the 3-digit industries. There is some evidence that stratification helps, but more so for high MSE estimates.

A last point is worth stressing. The weights adjustments considered here take place at the establishment level and so apply similarly to all fields reported by the establishment. This aspect allows for a single set of weights, and guarantees that subcomponents add up correctly to totals. For example, with field-specific weights, DAFW case totals plus DJTR case totals would not equal estimated totals for the quantity (DAFW cases + DJTR cases). But this aspect could as in table 4 leave us vulnerable to some induced bias.

The simulation results are rather hard-wired. Stratification is generally beneficial relative to simply taking the administrative records at face value. However, stratification need not eliminate all biases in the administrative records. Benchmarking the administrative records to SOII responses, referred to here as "calibration", may be helpful in estimating domains below the aggregate. Incorporating the external data is more helpful for smaller cells where SOII estimates are less precisely estimated.

4. Practical issues

The simulations above suggest there may be sensible and relatively manageable ways to incorporate administrative records into SOII estimation processes. The methods do not require record linkage, and do not appear to require information beyond that normally present in the SOII frame. However, there would certainly be practical challenges to overcome, and it is possible that record linkage or other information from past surveys or frames could improve estimation, and partially validate any chosen estimation method.

4.1 Issues surrounding post-stratification

The methods above generally require that we perform an analog to post-stratification on the administrative records. The control totals are essentially the number of business establishment units within strata. The SOII frame contains establishment counts but these counts are subject to updating. Some revisions occur directly at the time of SOII survey contact. For example, SOII staff may learn of a business closure only through efforts at mailing survey forms. A smaller survey component to the SOII would presumably make cell counts less accurate.

Administrative records would surely have errors complicating post-stratification. It is not clear that multi-establishment businesses would accurately separate reports by establishment, as required by the reporting

[12] I made the probability of reporting where there is a hospitalization slightly lower than average in some industries, so as to exacerbate the underinflation for such cases.

rules. OSHA recognizes this possibility and includes active language designed to prevent aggregated reports, but they have limited resources for audits or other data checks. Strictly speaking, the post-stratification strategies require disaggregation of a business entity in the same manner as it appears on the SOII frame.

Industry measurement error will occur in the administrative records. SOII processes use industry as it appears on the frame at the time of sampling, whereas OSHA records would contain self-reports of industry (please note, there are industry-based exemptions to reporting). It may be the case that industry measurement error is tolerable at say a 3-digit NAICS level but not at a detailed (6-digit) level, and that there are tradeoffs related to industry detail choice for the purposes of post-stratification of the OSHA source data.

4.2 Mode effects

The SOII asks respondents to report information as it exists on the forms that OSHA will require be submitted electronically. One might expect establishment reports to be similar under the two systems. Nevertheless, mode effects are possible. The SOII promises confidentiality whereas OSHA is a regulatory agency that intends to broadcast the required information via posting on a publicly accessible web site. It may be that different individuals report to the two systems, and at different times, or that SOII system processes such as re-contact or data editing result in data differences across systems. (Some SOII processes such as de-duplication of reports, handling of inconsistent reports, deletion of out-of-scope cases, and so forth would presumably be applied to the administrative data.)

It is also possible that the new rules induce establishments to alter their safety climate or their injury record-keeping processes in such a way as to change actual or reported outcomes as they apply to both data sources. (Changing actual outcomes is one implicit goal of the rules). This would not necessarily pose a problem for analysis or combination of current data but could make use of past information more difficult.

4.3 Useful exercises to conduct on OSHA data

The primary piece of information needed to combine the administrative records with SOII data is the probability that establishment *j* reports to the administrative system. The methods described above largely boil down to getting good measures for that probability, so as to undo reporting biases. Therefore it makes sense that a first step with actual data would be to determine how this probability varies with establishment characteristics and outcomes. For example, if the establishment DJTR case count is a predictor for reporting but DAFW cases are not, then a calibration exercise based on DAFW cases may not be the best strategy. Such an analysis would not require linked records.

However, record linkage would be required to settle other issues. Mode effects could be directly estimated. Linked records might also be used to determine whether multi-establishment firms are disaggregating along establishment lines as they exist in the SOII frame. Perhaps most importantly, direct comparisons of OSHA and SOII measures of establishment size and industry are needed to produce a mapping from OSHA data to the SOII strata. Lohr (2011) discusses a weights adjustment to unwind measurement differences for fields defining strata. Such an adjustment requires information on the particulars of the misclassification errors, which could in principle come from linked records. Any such adjustment could help with bias but inflate variances; realistic simulations might be informative.

4.4 Implications for sample allocations

Current SOII sample allocation sets sampling rates after accounting for certainty strata according to a Neyman allocation rule minimizing the variance of the topline incidence rate (Selby et al. 2008). This rule has a side benefit of being very productive for DAFW case records, so that sample sizes support comprehensive estimates of case and worker attributes. A substantial influx of administrative records

would seemingly force SOII to revisit the allocation process. First, SOII sample sizes might be cut. Second, allocation rules might target MSE measures and so incorporate bias concerns. One suggestion (Raghunathan 2015) is that scientific surveys' primary purpose may evolve toward providing information useful for triangulating to found data, and away from direct estimation. If so, SOII survey information in the OSHA overlap portion of the SOII population becomes more important, and this would seem to call for different objective functions for the allocation optimization problem.

5. Linking actual SOII and OSHA records

Although we do not yet have OSHA electronic reports under the new requirements, OSHA does post some information it receives under a different reporting requirement: OSHA receives reports on especially severe injuries, including amputations and hospitalizations.

I linked the 2015 OSHA data on severe injuries to 2015 SOII data in two ways. One linkage was an attempt to find OSHA-reporting establishments in the SOII frame. This linkage intends to determine feasibility of OSHA to SOII frame linkage, and also to quantify differences in industry coding in the two sources. A second linkage attempts to determine how often SOII DAFW cases on amputations and hospitalizations are reported to OSHA. This second linkage establishes that data are under-reported to OSHA for this particular set of reporting requirements. Readers are cautioned that record linkage as carried out here is a subjective process involving manual review, and further that under-reporting of severe injury cases can be expected to differ from under-reporting processes governing the new electronic submission rule.

OSHA requires all private sector employers to promptly report the circumstances surrounding certain severe incidents, including fatalities, amputations, and incidents resulting in hospitalization.[13] This reporting requirement extends to all industries and businesses of any size; the recordkeeping exemptions for OSHA forms 300, 300A and 301 described earlier do not apply for these severe incidents. Information reported includes company name, street address, city, state, zip code, a NAICS industry code, date of injury, and classification of injury circumstances based on narrative text. This information is potentially similar to a subset of information that would be required under the new electronic submission rules. However, the information does not include complete establishment reports for all cases, and other establishment level information (in particular, establishment employment).

5.1 Linkage to the SOII frame

An establishment reporting to OSHA should be in the SOII frame. I took OSHA reports from one state, Ohio, and attempted a record linkage to the SOII frame. The algorithm is:

1. Extract OSHA records for hospitalization cases occurring in Ohio during 2015. After removing duplicate records and dropping USPS records there are 491 records.
2. Extract private sector Ohio records for the 1st quarter 2015 from the BLS' Longitudinal Database (LDB, approximately equal to the SOII frame). Records include company legal name, trade name, and 3 addresses (2*3=6 combinations of name and address). There are approximately 275K establishment records.
3. Name/address fields in both sources are pre-processed with standardization software.
4. Probabilistic linkage. Linkage fields are name, street address, city, and 5-digit zip code. Industry is not used in linking records. Linkage is carried out 6 times, one for each LDB name/address

[13] OSHA receives these reports from employers operating in Federal-OSHA states. Some states operate their own job safety and health plans, under Federal OSHA approval. Severe injuries in those states do not appear in the Federal OSHA severe injury records.

combination, and possible links collated. A manual review of records was used to gauge how numerical match scores relate to probable matches.

Of the 491 establishment records from OSHA, 293 (about 60 percent) were clearly matched, in the sense that the OSHA record appeared in the LDB under nearly identical text for the linkage fields. All LDB name and address fields were useful in linking records. Among records that were clearly linked, there was a fairly substantial disagreement on industry codes in the two sources. The SOII and OSHA industry codes agreed approximately 58 percent of the time at the three-digit NAICS level and approximately 72 percent of the time at the 2-digit NAICS level.

Some qualitative information on linkage error is available. In addition to the 60 percent of records that appeared clearly matched, about 17 percent had a linked record that suggested a possible match, but with some divergence of representation in the two sources. These possible links might be resolved with access to additional data (for example, a federal tax identifier (EIN), or information on the entire case profile of the establishment). Some OSHA records were linked to multiple LDB identifiers; distinct LDB identifiers were found to have common company name and address information when looking at all name and address fields. This phenomenon appeared nonrandom with respect to industry. In addition, some LDB identifiers were linked to multiple OSHA records. Multiple linkage generally suggests a weak deduplication process.

Although these results come only from one experiment, they do suggest that record linkage at the establishment level would be difficult absent other information. They also suggest real challenges in conforming the administrative and SOII records along industry lines.

5.2 Linkage of hospitalization cases

Those SOII DAFW cases from federal-OSHA states that indicate an overnight hospitalization or amputation generally should be reported to OSHA. Therefore such a case found in SOII but not OSHA records can be presumed to reflect under-reporting to the administrative data, or linkage errors. A case found in OSHA records but not SOII may indicate a variety of situations (linkage error, SOII sampling rates less than one, a hospitalization that did not result in a day away from work, etc.). I took SOII case records for hospitalizations and attempted to find them in the OSHA records. This linkage is less informative than the attempt to link establishments, because one would not necessarily expect under-reporting for these case types to apply to the new recordkeeping rule to the same extent. Here is the algorithm:

1. Extract private sector 2015 SOII DAFW case records indicating overnight hospitalization, in federal-OSHA states. This results in 3799 records, with a weighted estimate of approximately 22,900 cases.
2. Extract 2015 OSHA hospitalization records, restricting to federal-OSHA states. This gives 7523 cases, about one-third of the SOII weighted total in federal-OSHA states.
3. Name and address fields in both sources are pre-processed with standardization software.
4. Probabilistic linkage. Linkage fields are name, street address, city, 5-digit zip code, state, and date of injury. Exact agreement on date of injury and state are required for an accepted link. Industry and case characteristics beyond identification as a hospitalization case are not used in linking records. Linkage is carried out 6 times, one for each LDB name/address combination, and possible links collated. A manual review of records was used to gauge how numerical match scores relate to probable matches.

Depending on how stringent one makes the linkage criteria, approximately 20-25% (unweighted) of eligible SOII hospitalizations appeared to show up in the OSHA records; weighted numbers are lower.

There are strong patterns in linkage probabilities along dimensions related to SOII cell strata. In particular, larger SOII units were much more likely to have their cases appear in the OSHA records.

Therefore we should not expect such substantial under-reporting with the new electronic reporting requirements (which apply to larger units). There were also differences across states and industries in linkage probabilities. In terms of industries, utilities and manufacturing establishments are much more likely to have their cases appear in OSHA records, while service and construction industry units were less likely to have their cases appear in OSHA records. Linked case records can show different industry codes in the different data sources; agreement at a three-digit NAICS level was comparable or slightly lower in these data to that described above for the establishment linkage.

6. Conclusions

This note describes various challenges that may confront BLS should it try to incorporate OSHA administrative records into its SOII estimation processes. Two general themes emerge. First, identifying the determinants of whether establishments report to OSHA or not is a key task. If the reporting determinants sufficiently relate to observables available on the SOII frame, stratification approaches hold some promise. Second, merging the two data sources through probabilistic record linkage could yield important information, particularly about the appropriate SOII stratum placement of OSHA-reporting establishments. Given that record linkage between SOII and actual OSHA data show substantial disagreement on industry, further investigation of estimator sensitivity to stratum misclassification appears warranted.

References

DiNardo, John, Nicole M. Fortin, Thomas Lemieux. 1996. Labor market institutions and the distribution of wages. Econometrica 64(5): 1001-1044.

Elliott, Michael R. and William W. Davis. 2005. Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. Journal of the Royal Statistical Society, Series C 54(3): 595-609.

Fay, R. and Herriot, R. 1979. Estimates of income for small places: an application of James–Stein procedures to Census data. Journal of the American Statistical Association, 74: 269–277.

Lohr, Sharon L. 2011. Alternative survey sample designs: sampling with multiple overlapping frames. Survey Methodology 37(2): 197-213.

Lohr, Sharon L. and J. Michael Brick. 2012. Blending domain estimates from two victimization surveys with possible bias. The Canadian Journal of Statistics 40(4): 676-696.

Raghunathan, Trivellore. 2015. Statistical challenges in combining information from big and small data sources. Working paper presented to NAS.

Särndal, Carl-Erik. 2007. The calibration approach in survey theory and practice. Survey Methodology 33(2): 99-119.

Selby, Philip N., Terry M. Burdette, and Erin M. Huband. 2008. Overview of the Survey of Occupational Injuries and Illnesses sample design and estimation methodology. Available at https://www.bls.gov/osmr/pdf/st080120.pdf .

Ybarra, Lynn M. R. and Sharon L. Lohr. 2008. Small area estimation when auxiliary information is measured with error. Biometrika 95(4): 919-931.

Appendix: Simulation parameters

Selection into reporting to OSHA. Establishments out of scope for the OSHA rules do not report to OSHA. Establishments in scope report with a probability assigned under one of three mechanisms. For purposes of the simulation, establishments with 50 or more employees that are in industries normally subject to (actual) OSHA oversight are assumed to be subject to the electronic reporting rule.

1. Selection on characteristics. Each stratum is assigned a random U[0,1] constant indicating the probability an establishment in that stratum reports.
2. Selection on total cases. Each establishment in the frame reports with probability p(z) = exp(z)/(1+exp(z)) where z depends on total cases as follows:

   z = 0 if total_cases<=1

   z = 0.6*ln(total_cases) if total_cases>1

3. Selection on case profiles. Probabilities of reporting depend on industry sector and whether an amputation or a hospitalization case occurs within the establishment. The assigned probabilities are largest for establishments in which an amputation case occurs. Reporting probabilities assigned are:

Industry Aggregate	Amputation present	Hospitalization but no amputation	Any other
Construction	0.8	0.6	0.7
Manufacturing	0.9	0.6	0.8
Natural Resources	0.8	0.7	0.7
Education, Health, Leisure and Hospitality	0.8	0.3	0.3
Financial Activities	1.0	0.5	0.5
Information	1.0	0.8	0.8
Professional, business and other services	0.9	0.7	0.7
Trade, transportation and utilities	0.7	0.5	0.5

Estimators.

Letting domain be indexed by *d*, establishments in SOII be indexed by *i*, and establishments reporting to OSHA be indexed by *j*, estimators are of the form

$$\hat{Y}_d = \theta_d \sum_{i \in SOII} I_{id} w_i y_i + (1 - \theta_d) \sum_{j \in OSHA} I_{jd} w_j y_j$$

where the I_d's are indicator functions for domain membership, the *w*'s are weights, and the *y*'s are case counts of various types. For all estimators the SOII weights w_i are sampling weights reflecting the inverse probability of selection.

1. SOII only, θ_d=1
2. OSHA only, θ_d=0 with weights w_j = 1.
3. OSHA and SOII equally weighted, θ_d=0.5 with OSHA weights w_j = 1.
4. OSHA and SOII weighted in proportion to domain-specific establishment counts from each source, $\theta_d=(n_{d,SOII} / (n_{d,SOII} + n_{d,OSHA}))$, with OSHA weights w_j = 1.
5. θ_d as in 4, with OSHA weights w_j post-stratified to SOII frame establishment totals N_d.
6. As in 5, with OSHA weights w_j post-stratified to SOII frame establishment totals N_d and additionally multiplied by a coefficient $\hat{\beta}$ from a regression

$$\hat{Y}_d^{SOII} = \alpha + \beta \hat{Y}_d^{OSHA} + \epsilon_d$$

where observations are at the 3-digit industry level and

$$\hat{Y}_d^{SOII} = \sum_{i \in SOII} I_{id} w_i y_i$$

$$\hat{Y}_d^{OSHA} = \sum_{j \in OSHA} I_{jd} w_j y_j$$

are estimates for total recordable cases reported in the industry to each source.

7. As in 5, with OSHA weights w_j post-stratified to SOII frame establishment totals N_d and then additionally multiplied by an adjustment factor from benchmarking to SOII weighted establishment counts, for groups determined by the establishment injury rate for total cases.

Estimators 1-4 are naïve incorporation of available OSHA data; results for 2-3 are dominated by 4 and are not reported. Estimator 5 is designed to overcome selection on characteristics. Estimators 6-7 are designed to reduce bias arising from selection on total cases.

Table 1: Percent of SOII Sample Subject to Electronic Reporting Requirements

	Establish-ments	Reported employment	Total cases	Cases with days away from work	Cases with days of job transfer or restriction
A. Unweighted					
No reporting	60.4	30.2	13.3	13.3	7.1
Summary records	28.0	16.0	22.0	24.0	23.4
Summary, case records	11.6	53.8	64.7	62.7	69.6
B. Weighted					
No reporting	92.3	58.2	35.0	36.6	20.8
Summary records	7.2	23.3	37.5	38.8	43.6
Summary, case records	0.6	18.5	27.6	24.6	35.5

Notes. Statistics sum to 100 percent within column for each panel. Estimates are based on 2015 SOII sample data.

Table 2: Root MSE as a Percent of True Values, Select Manufacturing Industries

Industry/Estimator	Total cases	FTE employment	DAFW cases	DJTR cases	DAFW hospital-izations	DAFW ampu-tations
Food Manufacturing						
SOII only	2.2	1.7	3.1	2.3	8.1	19.4
Naïve combination	15.1	14.5	15.4	15.3	15.9	22.7
Stratified weights	2.2	1.7	2.9	2.3	7.4	18.5
Beverage, Tobacco						
SOII only	5.5	3.0	7.6	7.6	36.3	47.4
Naïve combination	15.6	14.2	15.7	18.0	35.9	42.0
Stratified weights	4.7	2.4	5.8	6.0	35.2	41.7
Textile Mills						
SOII only	6.3	3.2	10.4	9.1	35.6	27.1
Naïve combination	13.2	12.5	13.0	15.2	35.8	22.1
Stratified weights	4.2	2.2	7.5	6.1	35.2	18.3
Textile Product Mills						
SOII only	8.9	3.9	13.5	13.8	43.7	45.6
Naïve combination	12.9	9.9	16.1	16.4	43.4	27.8
Stratified weights	8.1	3.1	12.9	12.3	41.8	23.7

Notes. All reported numbers are based on simulations. The probability an establishment reports to OSHA depends on observed establishment characteristics. See appendix for a description of data generation and the estimators.

Table 3: Root MSE as a Percent of True Values, Select Manufacturing Industries

Industry/Estimator	Total cases	FTE employment	DAFW cases	DJTR cases	DAFW hospital-izations	DAFW ampu-tations
Food Manufacturing						
SOII only	2.2	1.7	3.1	2.3	8.1	19.4
Stratified weights	5.3	2.3	5.5	5.4	8.0	18.8
Fay-Herriot	2.6	2.0	3.1	2.6	7.0	18.5
Calibration	1.9	1.7	2.6	1.9	7.0	18.5
Beverage, Tobacco						
SOII only	5.5	3.0	7.6	7.6	36.3	47.4
Stratified weights	7.2	3.1	7.9	8.7	35.2	41.2
Fay-Herriot	5.2	2.2	6.2	6.4	35.1	41.0
Calibration	4.5	2.2	5.6	5.5	35.2	40.9
Textile Mills						
SOII only	6.3	3.2	10.4	9.1	35.6	27.1
Stratified weights	7.2	2.6	9.0	8.9	35.3	18.5
Fay-Herriot	4.7	2.3	7.6	6.3	35.1	17.9
Calibration	5.0	3.8	7.6	6.8	35.2	18.3
Textile Product Mills						
SOII only	8.9	3.9	13.5	13.8	43.7	45.6
Stratified weights	9.9	3.4	13.5	14.0	42.2	18.1
Fay-Herriot	8.4	2.8	12.9	12.7	42.1	15.4
Calibration	8.3	4.0	12.9	12.5	42.2	18.5

Notes. All reported numbers are based on simulations. The probability an establishment reports to OSHA depends on the establishment's case total. See appendix for a description of data generation and the estimators.

Table 4: Root MSE as a Percent of True Values, US Private Sector Aggregate

Estimator	DAFW Hospitalizations	DAFW Amputations
SOII only	3.53	8.04
Stratified weights	3.80	10.14
Fay-Herriot	3.79	10.20
Calibration	3.84	10.01

Notes. All reported numbers are based on simulations. The probability an establishment reports to OSHA depends on whether the establishment experienced an amputation case. See appendix for a description of data generation and the estimators.

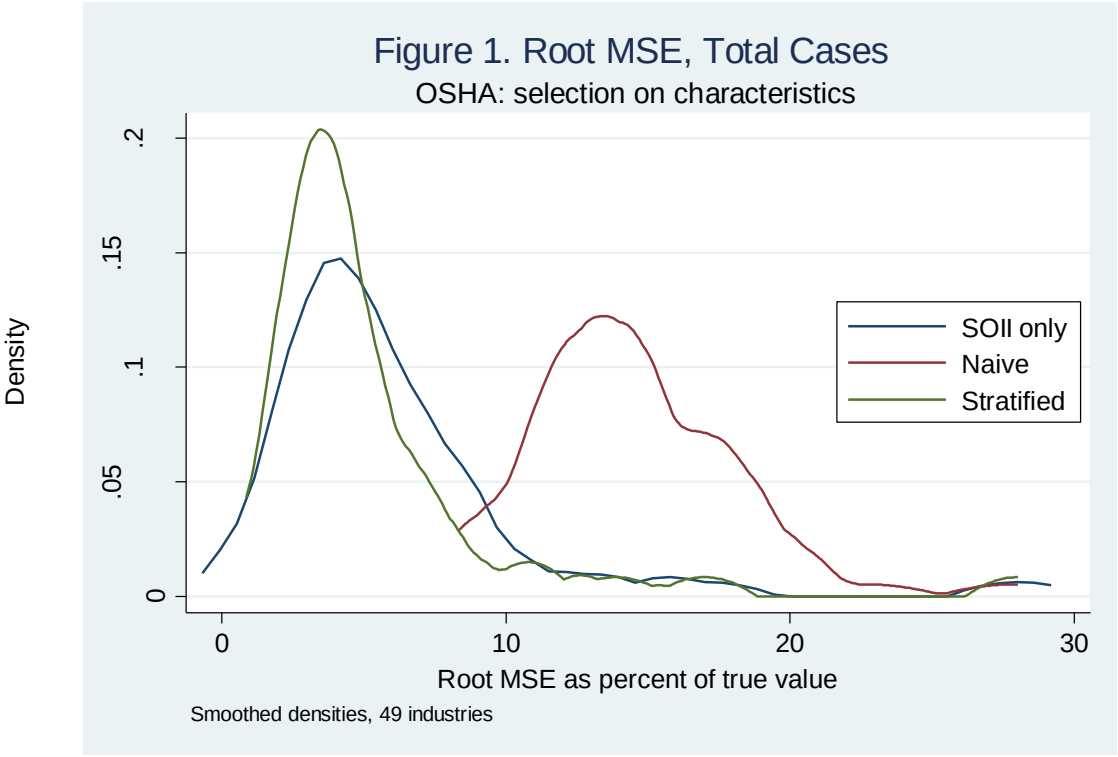
Figure 1. Root MSE, Total Cases
OSHA: selection on characteristics
Density
0
.05
.1
.15
.2
0
10
20
30
Root MSE as percent of true value
SOII only
Naive
Stratified
Smoothed densities, 49 industries

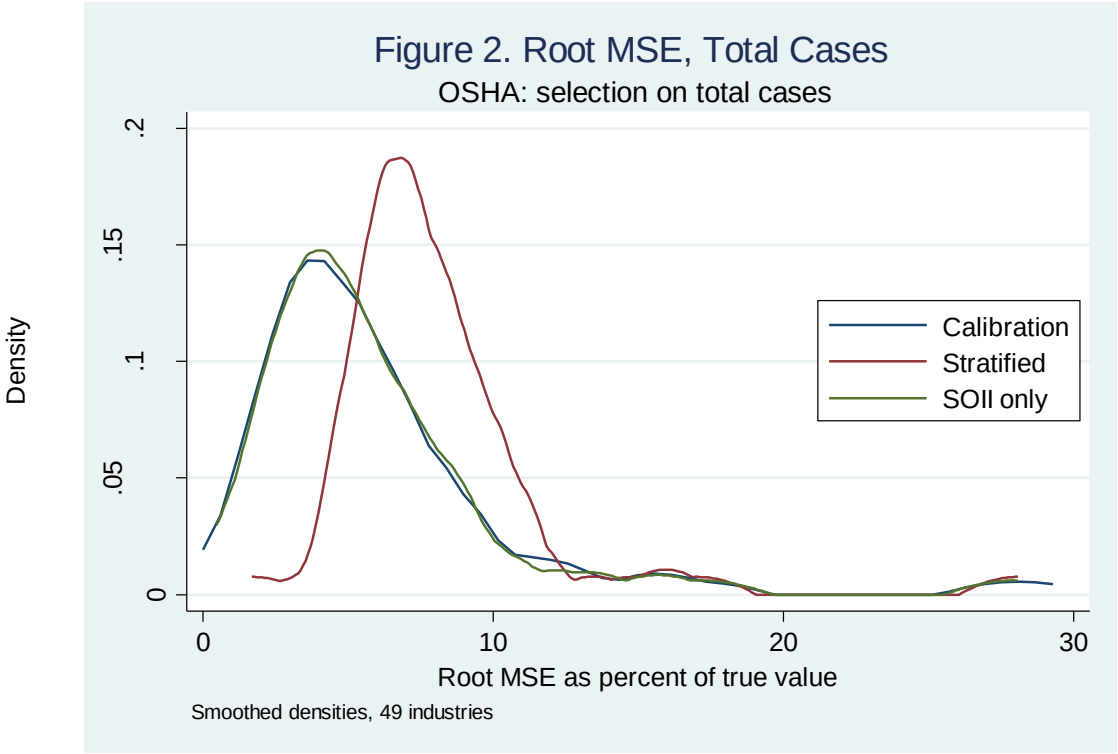
Figure 2. Root MSE, Total Cases
OSHA: selection on total cases
Density
0
.05
.1
.15
.2
0
10
20
30
Root MSE as percent of true value
Calibration
Stratified
SOII only
Smoothed densities, 49 industries

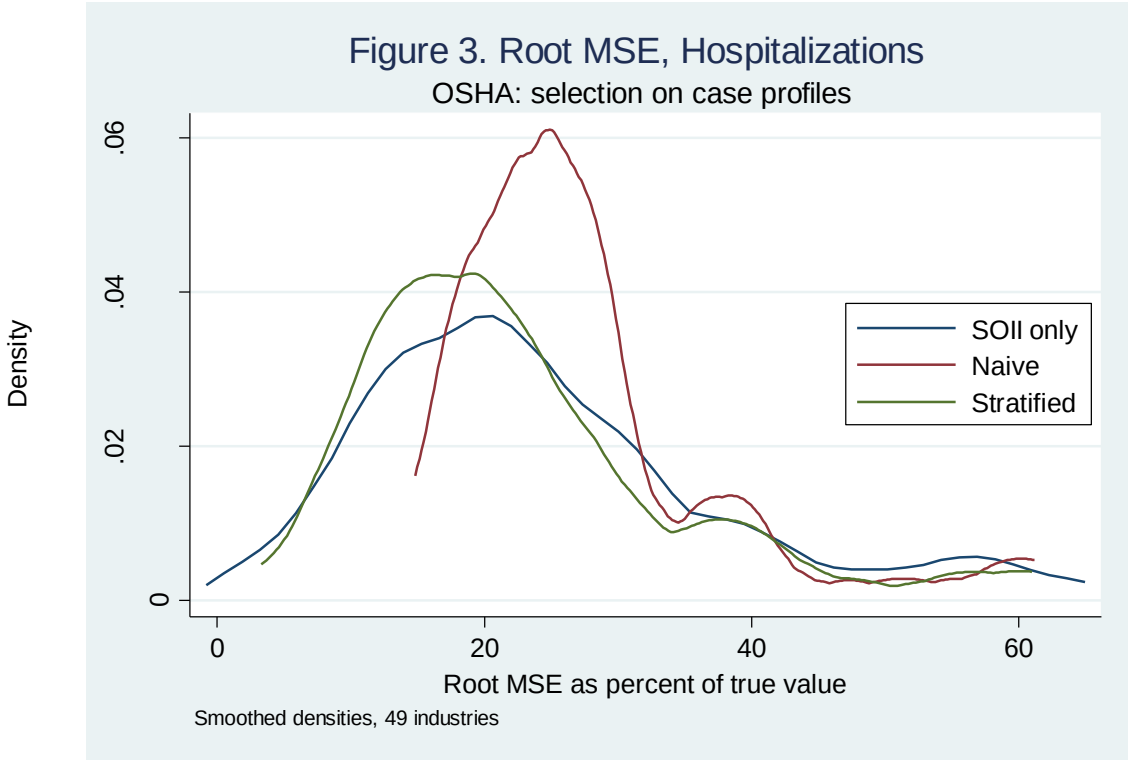
Figure 3. Root MSE, Hospitalizations
OSHA: selection on case profiles
Density
0
.02
.04
.06
0
20
40
60
Root MSE as percent of true value
SOII only
Naive
Stratified
Smoothed densities, 49 industries

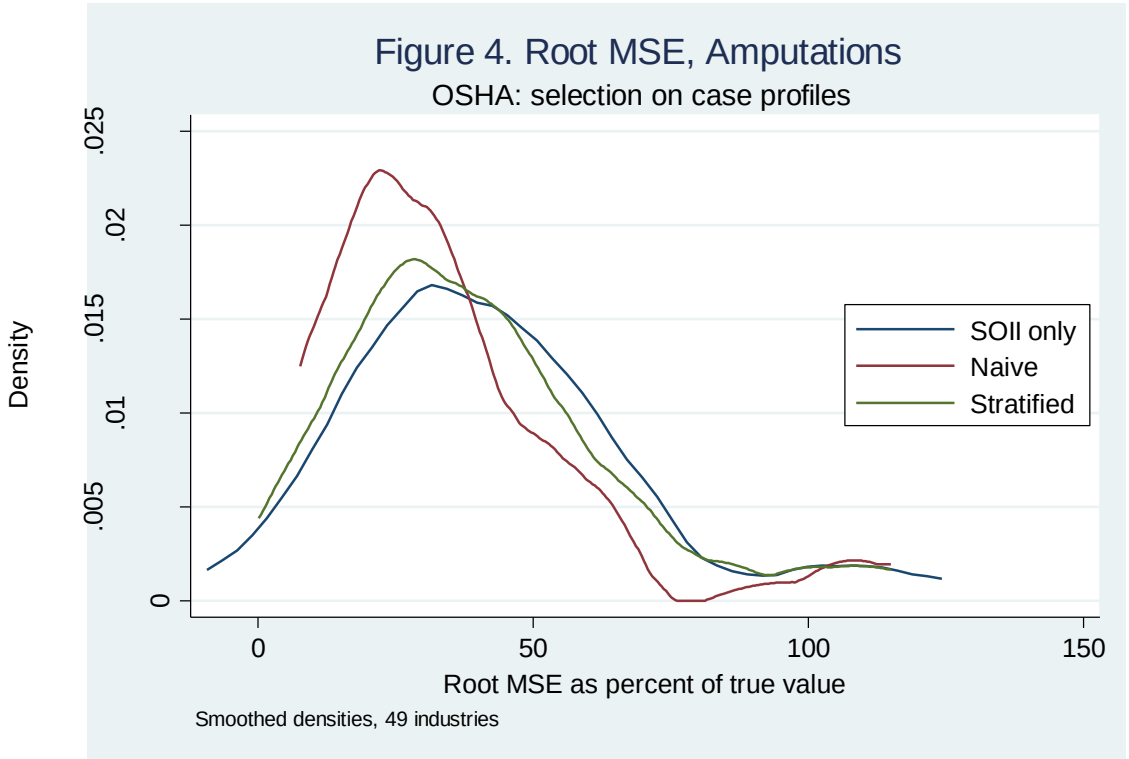
Figure 4. Root MSE, Amputations
OSHA: selection on case profiles
Density
0
.005
.01
.015
.02
.025
0
50
100
150
Root MSE as percent of true value
SOII only
Naive
Stratified
Smoothed densities, 49 industries