Comparing Injury Data from Administrative and Survey Sources: Methodological Issues October 2012

Nicole Nestoriak[1], Brooks Pierce[1]
[1]Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington DC 20212

Abstract

The Bureau of Labor Statistics' Survey of Occupational Injuries and Illnesses (SOII) is an important source of information for workplace injuries. Recent work comparing the SOII to Workers' Compensation (WC) administrative data concludes that the SOII undercounts injury and illness cases. Because the SOII is a sample, case by case comparisons between the two sources must distinguish between WC cases which are missing from the SOII because of sampling and cases which are missing because of underreporting. Previous research made this distinction by identifying SOII sampled establishments within WC data. This approach requires accurate employer information in WC data and subjective analysis of an establishment match. As an alternative, after matching SOII and WC data at the case level, we estimate the number of linked cases for the population by applying survey weights to the linked cases in the SOII sample. This allows us to estimate as a residual the number of WC cases missing in SOII, without having to directly identify SOII establishments in WC data. We describe the relative merits of this approach, and provide an alternative measure of the SOII undercount using Kentucky data.

Key Words: Administrative Data, Undercount

1. Introduction

There is growing evidence that the Bureau of Labor Statistics Survey of Occupational Injuries and Illnesses (SOII) substantially undercounts the true number of workplace incidents. Previous research attempting to quantify the extent of the undercount have produced a range of undercount estimates from 30% to 70%.[2] Because each of these studies varied not only in the state studied but also in methodology and source data, it is impossible to determine the cause of the large range of estimates. A more robust estimate of the undercount is the first step in determining whether and how BLS needs to alter its procedures in order to produce accurate workplace injury and illness statistics. This paper attempts to address these issues in hopes of producing a more robust methodology.

We define the undercount as occupational injuries and illnesses found in workers' compensation (WC) data but not in the SOII once exclusions have been made to the two data sources to have a common scope, or underlying population. Although the WC data may also be missing some workplace injuries and illnesses, the two data sources have

[2] See for example Boden and Ozonoff (2008), Leigh et al. (2004), Rosenman et al. (2006) and Ruser (2008).

much in common and choosing to compare SOII data to WC data allows for easy comparison with earlier research. The WC data used here, compiled from Kentucky first reports of injury, use a set of rules for inclusion similar to the SOII.[3] For each individual case, both data sources contain similar information: date of injury, date of birth, name of injured, nature of injury, etc. Despite these similarities, the data are collected for different purposes and by different means which make a direct comparison difficult. The SOII estimate of workplace injuries and illnesses is based on a sample of well-defined establishments, while the WC data are reports covering all employers in the state and contain some firm information. Additionally, there are differences in scope which are addressed below.

After identifying a common scope, one of the primary difficulties in comparing the SOII and WC is that the SOII is a sample. There are three potential strategies one could use to address the use of the SOII sample to make comparisons between the two data sources. The first option would be to look at case totals from both sources. The SOII data contain weights which allow one to compute a statistically valid estimate of the total number of injuries and illnesses. While this strategy is straightforward, it doesn't use all the detailed information available and it doesn't account for the possibility that each source is likely to miss some cases due to various reporting and measurement issues making the total undercount greater than the difference in case totals. A second option would be to match cases and establishments found in the two data sources. One could then restrict the set of WC cases to just those found in sampled SOII establishments. By utilizing the available details to do the case and establishment match, the output of this approach allows a detailed analysis of cases missed by one system or the other. The primary drawback of this approach is the limited information available on firms in WC to match to establishments in the SOII. Any errors in this match directly impact the estimated number of cases missed by the SOII.

A third potential approach, and the one focused on here, is a hybrid of the macro and micro approaches outlined above. One could use the detailed information to perform a case match but then use all of the WC cases in combination with information on the SOII sample to estimate the number of cases missed by the SOII. This approach avoids the difficult establishment match which relies strongly on the detail and quality of data on firms retained in WC systems. Because WC is a state specific program, this level of detail varies by state and prevents the creation of one uniformly appropriate strategy for matching WC claims to the appropriate establishment as defined in the SOII. While it is not possible to identify particular cases missed by SOII as in the micro approach above, the estimate of the undercount can be done separately by characteristics common to both datasets.

The following section of this paper provides details of the SOII and WC data relevant for measuring the undercount. The third section gives an overview of previous undercount research, followed by an overview of the steps necessary to measure the undercount in both the micro and hybrid approach. The fifth section provides a more detailed explanation of the hybrid methodology and a model to motivate its use. The sixth section

[3] Kentucky WC first reports in principle include all lost work time cases, as opposed to claims data or first reports from other states which only have lost work time cases that meet a waiting period criteria. However, this research strategy can easily be applied to other states, provided the WC recording of days of lost work time is reasonably conformable to SOII definitons.

contains results and the paper concludes with a comparison of the different methodologies and topics left for future research.

2. Details of the SOII and WC

The Survey of Occupational Injuries and Illnesses is a Federal-State Cooperative program. The mandatory survey is mailed to approximately 250,000 establishments each year. Employers provide information on workplace injuries and illnesses recorded on their Occupational Safety and Health Administration (OSHA) logs. Workplace injuries and illnesses in the Railroad and Mining industries are not collected using the survey but are provided by the Federal Railroad Administration and the Mine Safety and Health Administration respectively. In addition to the summary data providing total numbers of recordable cases, total days away from work (DAFW) cases, and total days of job transfer or restriction (DJTR) cases, additional worker demographics and case characteristics are collected for a subset of the DAFW cases, referred to as the case and demographics data. Additional details available for these cases include the name, date of birth, and sex of the injured worker and the date of incident, the nature of the disabling condition including the part of the body affected, and the event and source producing the condition.

The SOII is a stratified sample so that summary statistics can be published for a set of pre-determined industries for each state,. Each strata is defined by the establishment's state, industry, size class, and ownership (private, state government, or local government). The sample is allocated across strata in order to minimize variance in the total recordable cases incidence rate.[4] Each establishment is assigned a weight based on the sampling rate within its strata (some strata are certainty strata). This weight is then used in conjunction with the establishment level injury and illness information to estimate population level totals. Identifying the correct establishment is a key part of calculating a statistically valid estimate of these totals, and therefore, the BLS and its state partners take care in this process. Participants in the survey are notified of their selection prior to the beginning of the reference year. Data are then collected by BLS in the first few months of the year following the reference year.

The workers' compensation data cover all employers in the state of Kentucky with a date of injury in 2005. This data was compiled from first report of injury forms which are mandatory for all workplace injuries and illness with more than one day away from work. This rule for inclusion is similar to that for the SOII case and demographics data of at least one day away from work beyond the day of injury. For workers, the WC data include the worker's name, date of birth and gender. For injuries and illnesses, the data include the date of injury, part of body and nature. The disadvantage of using the first report of injury data is the limited information on the final disposition of the case.

As WC is an insurance program designed to provide wage replacement and medical benefits to injured workers, there may be records included in these files that would not be covered by the SOII and records included in the SOII that would not be included in WC. Examples of the first type of record include late developing cases in which the worker is injured in 2005 but does not require time away from work until 2006 or later. Other possibilities include cases which were initially disputed by the employer for either not being work related or not being serious enough to require time away from work, but for

[4] See U.S. Department of Labor (2012) and Selby, Burdette, and Huband (2008) for details.

which the employer ultimately maintained responsibility through the WC system. Examples of the latter situation include cases with too few days away from work to merit indemnity benefits, so that the worker does not notify the employer of the injury or file a WC claim, and therefore no first report of injury was made.

Another issue with using WC data to compare to the SOII data is the limited information on the employer. While the first report form asks for a company name, address, industry, and Federal Employer Identification Number (EIN), companies may choose to report the information in a variety of ways. An employer can record a physical address or a mailing address. The EIN may be that associated with the particular establishment or may be that of a parent company. A company name may reflect a trade name or a legal entity. These variations make direct matches of employers between WC and the SOII difficult.

3. Previous Undercount Research

A number of studies have attempted to estimate the size of the SOII undercount. Leigh, Marcin, and Miller (2004) reference a variety of earlier macro studies and estimate the undercount to be between 30 and 70% of injuries in 1999. Oleinik and Zaidman (2004) compared case totals with 4 or more days away from work in the Minnesota SOII and WC data and found a level of concordance above 90%. Rosenman et al. (2006) examined the SOII, WC data, OSHA Integrated Management Information System, and Occupational Disease Reports for Michigan for the years 1999, 2000, and 2001 and found an undercount between 60 and 70%. Boden and Ozonoff (2008) perform a similar exercise to that of Rosenman et al. (2006) for an additional six states and find an undercount with a range between 30 and 45%.

Building on the earlier work of Boden and Ozonoff (2008), Nestoriak and Pierce (2009) utilize the matched SOII-WC dataset that was an output of their work for the state of WI in order to determine what person, employer, and injury characteristics make an injury more or less likely to be reported to SOII. They find that cases that are severe and have sudden onset, such as an amputation, have higher SOII capture rates than cases that become apparent over time, such as carpal tunnel syndrome. Additionally, cases filed with WC after the reference year have a lower capture rate. Finally, employers with multiple establishments in the same state have lower than average capture rates which is likely the result of methodological issues in linking SOII establishments to WC employers.

4. Steps to Measure the Undercount

In both the micro approach followed by Boden and Ozonoff (2008) (hereafter, Boden-Ozonoff), and the hybrid approach outlined below, the first step in measuring the undercount is to match SOII cases and WC cases using the available detail on worker's name, sex, date of birth, date of injury, the nature of the injury, and some employer information. While a sizable fraction of cases can be matched deterministically, a higher overall match rate can be obtained by employing additional matching strategies. Probabilistic matching loosens the criteria that all of the fields must match and places greater weights on fields that match with unique values. String and numeric comparators

make allowances for typos while determining if fields agree. Details of the case match for Kentucky can be found in the results section.

After matching, adjustments are made to the data so that each data set refers to a common underlying population, or scope. Cases are often dropped from one or both datasets so that the set of injuries and illnesses cover a common set of industries and rule for inclusion, often defined by the number of days away from work. Following these restrictions, there are three types of cases: in the SOII but not WC, matched SOII-WC, and in WC but not the SOII. This final set of cases is not equal to the SOII undercount due to the sampling used in the SOII. While some of the cases found only in WC are likely at sampled establishments and therefore part of the undercount, another set of these cases are outside of the SOII sample. Cases not sampled by the SOII are still reflected in the SOII estimated totals once one applies the sample weights. Determining how to divide the WC cases not found in the SOII into a subset that is in sampled establishments versus a subset that is not differentiates the micro and hybrid methodologies.

In the micro approach, the goal is to keep only the WC claims in SOII sampled establishments which therefore requires determining which WC employers were sampled by SOII. There are potentially three ways one could accomplish this match. The first would be to use the already matched SOII-WC cases and link the SOII employer with the WC employer. An alternative would be to do an employer match similar to the case match using the company name and address information. Finally, many states have an EIN on their WC claims which can be matched directly to the SOII. Once it is determined that a company in WC matches an establishment in the SOII, a WC company identifier (often the EIN) must be used to determine the set of WC cases associated with the sampled establishment.

After matching employer information, further adjustments are necessary to account for the differences in how company information is recorded in the SOII and WC. In the SOII, the sampling unit is the establishment which is a single physical location of a firm. In WC, companies may report establishment or more aggregated firm information. If the EIN has one establishment in the state, all cases are kept. If the EIN has more than one establishment in the state, the next step is to determine how to account for having only a fraction of a WC company sampled by the SOII. Boden-Ozonoff accomplished this by calculating the fraction of employment in sampled establishments versus the full EIN employment for each multiple establishment EIN in which there was a WC-SOII match. Each of the affected WC cases was then down-weighted so that the weighted total of WC cases divided by the total number of WC cases in the EIN equaled the sampled employment divided by the full employment at the EIN. This adjustment implicitly assumes that the sampled and unsampled portions of a given multiple-establishment firm have similar injury rates. Although there is little empirical evidence on whether or not this assumption is valid, we feel it is unlikely to have much impact on the final results.

The previous two steps are key to defining the undercount as they determine the set of WC cases that were at sampled SOII establishments but not found in the SOII. False matches or missed matches can move sets of WC claims between the in-sample or not-in-sample categories.

In attempting these steps using the Kenutcky WC data, different strategies for matching companies between WC and SOII yielded different results. Additionally, the EIN information in the Kentucky WC data was often inconsistent with BLS data. Matched

WC-SOII cases sometimes had different EINs in WC and the SOII. Approximately 20% of the EINs in the Kentucky WC could not be found in the BLS universe files from which the SOII sample is drawn. These inconsistencies led us to pursue an alternate, hybrid methodology for measuring the SOII undercount.

In the hybrid approach, the goal is to look at the three types of cases defined after matching cases, and apply the SOII sampling weights to determine what fraction of them represent the undercount. In particular, cases in WC but not the SOII sample fall into one of three categories. The first category includes cases in sampled SOII establishments; these cases are part of the undercount. The second category includes cases that are not in sampled SOII establishments, but would have been linked to cases in the SOII if the SOII had surveyed all establishments. These cases are not part of the undercount. The third category includes cases that are not in sampled SOII establishments, and would not have been linked to the SOII even if the SOII were a census. These cases are part of the undercount. The hybrid approach uses the SOII weights to estimate the number of WC cases that would not match to the SOII even if the SOII were a Census.

5. Hybrid Method

As mentioned above, the hybrid method begins, just as the micro approach, by matching cases in the SOII to cases in WC using detailed case information. After matching, each of the SOII cases is retained with an additional characteristic: has the case also been matched to a WC case? The WC cases are collected from all companies within a state and are therefore treated as a census of cases. If the SOII were also a census, calculating the number of cases in both or one source only would be straightforward. However, because the SOII is based on a sample, cases that are not matched in WC should not necessarily be considered missed by the SOII.

Applying the SOII sampling weights to the linked SOII-WC cases gives an estimate for the population of matched SOII-WC cases. One can then use this estimate to net out the matched cases from the WC totals to get an estimate for the WC only cases, or cases missed by the SOII, for the population.

Table 1: Hybrid Approach

	SOII only ($I^{SOII\sim}$)	Matched SOII-WC (I^{M})	WC Only ($I^{WC\sim}$)
Sample (I_S)	$I_S^{SOII\sim}$	I_S^{M}	
Not in Sample (I_N)			$I_N^{M} + I_S^{WC\sim} + I_N^{WC\sim}$

The above table illustrates the hybrid approach and highlights the two dimensions on which one must consider the data. The rows or subscripts differentiate between cases associated with sampled establishments and cases in establishments that were not sampled. The columns or superscripts differentiate between cases that are in the SOII only (SOII~), matched SOII-WC (M), or WC only (WC~). Before matching we have a set of SOII cases, I^{SOII}, and a set of WC cases, I^{WC}. After the case match, there are three groups of cases: in the SOII sample but not in WC, linked SOII-WC, and in WC but not the SOII sample. The final group of cases is not equivalent to the undercount because it includes cases in establishments not sampled by the SOII that would have been SOII-WC matches had the BLS surveyed all establishments. We can estimate this group of injuries,

I_N^M, by applying SOII sampling weights to the matched SOII-WC cases in sampled establishments, I_S^M. Letting j index SOII cases, with sampling weights w_j^{SOII} (where $j \in I_S^M$ indicates the cases in which there was a match between the SOII and WC), one can estimate the number of WC-only cases which are the undercount, as

$$\hat{I}^{WC\sim} = I^{WC} - \sum_{j \in I_S^M} w_j^{SOII} = \left(I_S^M + \left[I_N^M + I_S^{WC\sim} + I_N^{WC\sim}\right]\right) - \sum_{j \in I_S^M} w_j^{SOII}$$

This estimates the matched SOII-WC for both the sample and the establishments not in the sample combined as the residual left after subtracting the weighted SOII matched cases from the WC total, I^{WC}. From these totals it is possible to calculate the SOII capture rate as defined by

$$\widehat{SOII\ Capture\ Rate} \equiv \frac{\hat{I}^{SOII\sim} + \hat{I}^{M}}{\hat{I}^{SOII\sim} + \hat{I}^{M} + \hat{I}^{WC\sim}}$$

To make things more concrete, consider a model in which workers are either injured or not, and injured workers choose whether to report their injury to any of two sources. Both injuries and reporting are random variables. For simplicity, we first assume the sources (call them A and B) are censuses, although the assumption will be relaxed below.

Workers $\quad j = 1,2,\ldots,N$

Injuries $\quad i_j^* = \begin{cases} 1 & if\ injury\ to\ j \\ 0 & if\ no\ injury\ to\ j \end{cases}$

Reporting behavior, should the worker be injured

$$R_j^A = \begin{cases} 1 & would\ report\ to\ source\ A \\ 0 & would\ not\ report\ to\ source\ A \end{cases} \quad j = 1,2,\ldots,N$$

$$R_j^B = \begin{cases} 1 & would\ report\ to\ source\ B \\ 0 & would\ not\ report\ to\ source\ B \end{cases} \quad j = 1,2,\ldots,N$$

Here reporting behavior is conditional on an injury ($i_j^*=1$), and therefore we assume there are no instances of over-reporting. Injuries and reporting behaviors are unobservable. What one observes in the data is the product of the two.

$$i_j^A = R_j^A i_j^* = \begin{cases} 1 & injury\ reported\ to\ source\ A \\ 0 & no\ report\ to\ source\ A \end{cases}$$

$$i_j^B = R_j^B i_j^* = \begin{cases} 1 & injury\ reported\ to\ source\ B \\ 0 & no\ report\ to\ source\ B \end{cases}$$

Therefore $i_j^k=0$ could mean either no injury or an injury that goes unreported to source k (but $i_j^k=1$ implies an actual injury as there are no false reports). $E(R_j^k)$ can vary by j and k, so different people may have different reporting propensities, and the same person may have different reporting propensities to different sources. For a given person, reporting behavior need not be statistically independent, so the covariance between R_j^A and R_j^B need not equal zero.

While we observe individuals' reported injuries, we are interested in injury totals, reported population totals for each source, and capture rates. Specifically, we would like estimates for

$$Total\ cases = I^* \equiv \sum_{j=1}^{N} i_j^*$$

$$Cases\ reported\ to\ A = I^A \equiv \sum_{j=1}^{N} i_j^A$$

$$Cases\ reported\ to\ B = I^B \equiv \sum_{i=1}^{N} i_j^B$$

$$Source\ A\ total\ capture\ rate = \frac{I^A}{I^*}$$

$$Source\ B\ total\ capture\ rate = \frac{I^B}{I^*}$$

While we would like to ultimately have capture rates defined as above, reported cases for each source divided by total cases, the I* are not directly observable, so as an interim step we are interested in capture rates among reported cases.

In order to calculate these rates, we first define the total number of unique cases reported to any source. This total is the sum of cases found in both systems, cases found in source A only, and cases found in source B only. Using the notation above, a case is in both systems if $i_j^A i_j^B$=1. A case is in source A only if $i_j^A(1 - i_j^B)$=1 , and analogously for source B. The unique case total is

$$\begin{aligned} I &\equiv \textstyle\sum_{j=1}^{N} \max\,(i_j^A, i_j^B) \\ &= \textstyle\sum_{j=1}^{N} i_j^A i_j^B + \sum_{j=1}^{N} i_j^A(1 - i_j^B) + \sum_{j=1}^{N}(1 - i_j^A) i_j^B \\ &= \textstyle\sum_{j=1}^{N} i_j^A + \sum_{j=1}^{N}(1 - i_j^A) i_j^B \qquad (1) \\ &= \textstyle\sum_{j=1}^{N} i_j^A + \left\{\sum_{j=1}^{N} i_j^B - \sum_{j=1}^{N} i_j^A i_j^B\right\} \\ &= I^A + \{I^B - I^{AB}\} \end{aligned}$$

The final equation (1) will be useful below in estimating each of these quantities. Consistent with the above equations, one can also define the capture rate among the observable cases.

$$Source\ A\ reported\ capture\ rate = \frac{I^A}{I}$$

$$Source\ B\ reported\ capture\ rate = \frac{I^B}{I}$$

Equation one assumes that both sources are censuses as both summations are over 1…N. In order to incorporate sampling in source A, as there is in the SOII, imagine we observe i_j^B for j=1,2,…N, but that we only observe i_j^A for j=1,2,…n. The remaining i_j^A for j=n+1,…N are reported cases outside of the sample and are not observable. They are the cases that would have been reported had source A been a census instead of a sample. The SOII is a stratified random sample of establishments, with sampling fractions f and hence weights w=(1/f) that vary by strata.

Returning to equation 1, we would like estimators for the first and third terms. Using the SOII sampling weights, we can construct

$$\hat{I}^A = \sum_{j=1}^{n} w_j i_j^A$$

$$\hat{I}^{AB} = \sum_{j=1}^{n} w_j i_j^A i_j^B$$

$$\hat{I} = \hat{I}^A + \left\{ I^B - \hat{I}^{AB} \right\}$$

and use the resulting estimates to calculate the source-specific capture rates among reported cases. This procedure is the traditional Peterson estimator from the capture-recapture literature, as adapted for sampling considerations, except that we do not estimate the number of cases that go unreported to any list.[5] That is, we estimate capture rates using $\hat{I}$ rather than I^* in the denominator. We do this not because it is logically preferred (in fact, one would prefer to have in hand a good estimate for I^*), but rather because we want to focus on the problem of obtaining good estimates for I in our particular application. Absent matching errors and SOII non-sampling error these estimators are unbiased. Injury propensities, reporting propensities and reporting covariances can differ by strata, or by worker, without biasing the resulting estimates. Also note that calculating capture rates over the whole population does not require that one know the strata of source B observations. Furthermore, this exercise can be repeated by sample strata so long as one knows to which strata the source B observations belong.

More generally, one can perform the above exercise to calculate capture rates by any data characteristic which is commonly defined in both data sources. An example, which is shown below, is to calculate the capture rate by month of injury. Other potential characteristics by which one might hope to calculate capture rates include nature of injury and single versus multiple establishment firms. However, both of these characteristics are likely to be defined differently in the different data sources, and therefore capture rates calculated using the hybrid approach by nature of injury or type of establishment are not likely to be accurate without further adjustment. For example, it may be difficult to estimate capture rates for amputations using these methods if such injuries are often recorded as crushing injuries in one source. In addition to concerns in defining a common scope for more narrowly defined groups, the more general assumptions necessary for the hybrid approach must also hold for each group. For example, the propensity to report July injuries must not systematically differ between the sample and the population. Similar

[5] See Wolter (1986) and references therein for derivations and discussion of similar estimators with application to coverage estimates of Census data.

considerations apply to the propensity to report July injuries to source A, conditional on reporting to source B.

In addition to the assumptions outlined above, accuracy of the hybrid methodology estimates is limited by the accuracy of case matching and creation of a common scope. These first two steps in the hybrid approach (and the micro approach) define the set of linked and unlinked cases. Errors in defining cases which are in scope will affect the overall number of cases while errors in matching will shift cases between the linked and unlinked categories. These types of errors will have a direct impact on the estimate of the undercount regardless of the methodology used. Because of the differences across states and subjective nature of matching, it is difficult to quantify the impact of these types of errors on the final estimates. Additionally, the WC data are continuously updated with determinations as to work relatedness while the SOII captures a snapshot of injuries and illnesses shortly following the reference year. Case matching depends upon the accuracy of recording name and date of injury, which are of limited use in the SOII (and therefore not checked for accuracy by BLS) but a key factor for determining the amount of benefits in WC. Despite these difficulties, comparisons of the SOII and WC cases are perhaps the best approach to ascertain the extent of a potential SOII undercount, and the hybrid methodology is a potentially useful complement to deriving undercount estimates on an establishment by establishment basis.

6. Results

The SOII data used for this estimate of the undercount is for the state of Kentucky in 2005. In order to match cases to WC, the case and demographics data was used. The workers' compensation data was created from an extract of the Kentucky first reports of injury for injury dates occurring in 2005. The extract was created in July of 2008.

Table 2: Case Totals

	SOII	SOII Weighted (SE)	WC
Overall	5,086	25,490 (1,230)	33,540
After exclusions	4,333	24,560 (1,229)	30,525

The SOII case and demographics data has 5,086 cases for Kentucky in 2005. Cases in which the industry was Railroad or Mining were removed from the dataset because the data for these industries is not collected using the survey and do not have detailed person information necessary for matching. Cases from the Temporary Help Services industry were also removed as reporting requirements for this industry are different under OSHA and WC rules. After exclusions there are 4,333 cases which yield a weighted estimate of 24,560 cases in Kentucky. Standard errors reflect sampling variability. A similar set of exclusions were made to the WC data and the final case count there was 30,525. Unlike WC data collected by other states, the Kentucky WC first reports contain all cases with at least one day away from work, therefore no further restrictions on the SOII by number of days away from work are necessary.

The cases were matched in an iterative process. The first step involved a match of 1,586 cases with an exact match on certain person, injury and establishment characteristics. The second step took the residuals from step one, required an exact match on the EIN and

allowed for probabilistic matches on person and injury characteristics for 750 more matches.[6] The third step again took the residuals from the previous step, but was less restrictive in that no field required an exact match but again person, injury and establishment characteristics were used in a probabilistic match which yielded 456 additional matches. Two further sets of probabilistic matches added additional fields for matching, and used date of birth and date of injury as exact match fields, respectively. These two sets of matches yielded 22 more cases combined. Probabilistic matches were reviewed by three people and majority rule determined final match status. A final deterministic rule declared a case matched if it occurred within a linked establishment (the WC firm and SOII establishment shared an already-linked case), and the two sources agreed on date of injury, employee name and age within one year; this rule found an additional 10 matched cases. The case totals for the number of matched and unmatched cases by source are in Table 3.

Table 3: Case Matches

	SOII only	Matched SOII-WC	WC Only
Sample	1,509	2,824	
Not in Sample			27,701

Calculating the SOII capture rate then requires applying weights to the matched, in-sample cases which yields 16,030 cases (se=818),[7] and then finding the set of in WC and not in SOII cases. The universe of cases is determined by adding these two sets of cases with a weighted total of SOII only. The final SOII capture rate is 62.9% as shown below. This result falls within the range found using the micro methodology for other states.

$$\hat{I}^{WC\sim} = \left(I_S^M + \left[I_N^M + I_S^{WC\sim} + I_N^{WC\sim}\right]\right) - \sum_{j \in I_S^M} w_j^{SOII}$$

$$= (2{,}824 + [27{,}701]) - 16{,}030 = 14{,}495$$

$$\widehat{SOII\ Capture\ Rate} = \frac{\hat{I}^{SOII\sim} + \hat{I}^{M}}{\hat{I}^{SOII\sim} + \hat{I}^{M} + \hat{I}^{WC\sim}} = 62.9\%\ (\text{SE}=02.47)$$

As mentioned above, the SOII capture rate can be calculated separately by any characteristic which is found in both datasets. Table 4 shows the SOII capture rate by month of injury. As was found in previous work,[8] the SOII capture rate appears to be lower at the end of the year. One hypothesis for this finding is that employers are late in recording injuries on the OSHA log before the SOII data are collected early in the following year. Other patterns in the monthly capture rate can possibly be explained by the composition of injuries in combination with reporting behavior, a topic left for further research.

[6] Probabilistic matching used LinkPlus software, available from the Center for Disease Control, which is based on Fellegi and Sunter (1969) and Belin and Rubin (1995).
[7] Details on calculating standard errors for the linked cases and the SOII capture rate are provided in an Appendix.
[8] See Nestoriak and Pierce (2009).

Table 4: SOII Capture Rate, by Month

Month	SOII capture rate (SE)	Month	SOII capture rate (SE)
January	66.92% (4.08)	July	68.72% (4.02)
February	61.42% (3.57)	August	65.25% (3.84)
March	57.13% (3.80)	September	67.73% (4.48)
April	64.30% (4.16)	October	59.91% (3.63)
May	65.73% (3.99)	November	57.52% (3.77)
June	62.23% (3.49)	December	55.02% (3.16)

While it is not possible to directly compare results from the hybrid and micro approach using Kentucky data, as a robust establishment match was not possible, a separate comparison was made using the Boden-Ozonoff Wisconsin data from 1998 through 2001. These results are calculated from the same set of case matches.[9] Comparing the two sets of results in table 5 below, the hybrid approach yields modestly higher capture rates. However, it is likely not appropriate to extrapolate the Wisconsin results to other states as Wisconsin had higher than average quality firm data in their WC extract.[10]

Table 5: SOII Capture Rate, Comparing Different Approaches

Wisconsin SOII Capture Rate, 1998-2001	
Micro (Boden-Ozonoff)	Hybrid
70.0%	73.8%

7. Conclusion

While this paper has focused on the hybrid approach, it is not necessarily the preferred methodology for all scenarios. If one were trying to measure the SOII undercount with WC data that had good information on firms, the micro approach has some advantages. With the micro approach, one can examine case by case the types of cases that are more likely to be missed by the SOII. In addition to examining SOII capture rates by characteristic, it is possible to do a multivariate analysis looking at all the case characteristics within one analysis and therefore examining the impact of one characteristic holding all of the others constant. More detailed analysis of the undercount by strata or by case characteristics might suggest particular ways to improve the SOII.

The advantages of the hybrid approach are strongest when the WC data does not have good information on firms. One example is when the only common identifier is the EIN which is not recorded consistently. While a full multivariate analysis is not possible, the SOII capture rates can be calculated by characteristic. The hybrid approach may also be preferred when comparing results across states. While differences in the quality of the case data may affect the quality of the case match, differences in the quality of the firm

[9] Les Boden, with an agreement from the state of Wisconsin, kindly provided the results of his case match and all of his accompanying programs to the authors. Results reported here are based on table 2 of Boden-Ozonoff; in later tables they apply adjustments for cases missed by both sources.

[10] The Wisconsin data have Unemployment Insurance (UI) account numbers for all cases. The UI number appears to better facilitate matching SOII and WC data than does the EIN.

data will have less of an impact on the final SOII capture rate. Further, the hybrid approach removes some of the subjectivity in matching, making for a more easily replicable methodology.

Acknowledgements

We would like to thank John Ruser for the original idea which motivates this paper and Anthony Barkume, Gwyn Ferguson, Jeffrey Gonzalez and participants at the 2011 National Occupational Injury Research Symposium for comments. Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

References

Belin TR, Rubin DB (1995). A method for calculating false-match rates in record linkage. *Journal of the American Statistical Association* 90(430): 694-707.

Boden LI, Ozonoff A (2008). Capture-recapture estimates of nonfatal workplace injuries and illnesses. *Annals of Epidemiology*. 18(6): 500-506.

Fellegi IP, Sunter AB (1969). A theory for record linkage. *Journal of the America Statistical Association* 64(328):1183-1210.

Leigh JP, Marcin JP, Miller TR (2004). An estimate of the U.S. Government's undercount of nonfatal occupational injuries. *Journal of Occupational and Environmental Medicine.* 46(1): 10-18.

Nestoriak N, Pierce B (2009). Comparing workers' compensation claims with establishments' responses to the SOII. *Monthly Labor Review*, May: 57-64.

Oleinick A, Zaidman B (2004). Methodologic issues in the use of workers' compensation databases for the study of work injuries with days away from work. I. Sensitivity of case ascertainment. *American Journal of Industrial Medicine,* 45: 260-274.

Rosenman KD, Kalush A, Reilly MJ, Gardiner JC, Reeves M, Luo Z (2006). How much work-related injury and illness is missed by the current national surveillance system? *Journal of Occupational and Environmental Medicine*, 48(4): 357-365.

Ruser, JW (2008). Examining evidence on whether BLS undercounts workplaces injuries and illnesses. *Monthly Labor Review*, August: 20-32.

Selby PN, Burdette TM, Huband E (2008). Overview of the Survey of Occupational Injuries and Illnesses sample design and estimation methodology. In *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association. 1337-1344.

U.S. Department of Labor, Bureau of Labor Statistics (2012). Handbook of Methods. Accessed at http://www.bls.gov/opub/hom/pdf/homch9.pdf.

Wolter, KM (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81: 338-346.

Appendix: SOII Case & Demographics Variance Estimates

Here we describe how we calculate variance estimates for the number of linked and unlinked SOII cases, and the SOII capture rate. Let

I^M = weighted total for linked SOII cases
$I^{SOII\sim}$ = weighted total for unlinked SOII cases
$I^{SOII} = I^M + I^{SOII\sim}$ = weighted total for all SOII cases , and
$p = I^M / I^{SOII}$ = proportion of SOII cases that are linked.

We wish to estimate the variance $V(I^M)$. Assuming that p and I^{SOII} are independent,

$$V(I^M) = V(p\, I^{SOII}) = (I^{SOII})^2 V(p) + p^2 V(I^{SOII}) + V(I^{SOII})\, V(p),$$

from the formula for the variance of the product of two independent variables. Hence estimates for p, I^{SOII}, V(p) and $V(I^{SOII})$ suffice to form an estimate for $V(I^M)$. Estimates for I^{SOII} and $V(I^{SOII})$ come directly from published private sector totals, adjusted for the excluded sectors. V(p) is the usual variance of a proportion.[11] An analogous calculation gives the variance of unlinked SOII cases, $V(I^{SOII\sim})$.

The SOII capture rate we use in the text is defined as

$$\text{SOII capture rate} = (I^M + I^{SOII\sim}) / (I^M + I^{SOII\sim} + I^{WC\sim}),$$

where $I^{WC\sim}$ is the number of WC-unlinked cases. We estimate the variance of the SOII capture rate using a first order Taylor series approximation, treating I^M and $I^{SOII\sim}$ as subject to sampling error. This requires the estimates $V(I^{SOII\sim})$ and $V(I^M)$ shown above, and an estimate for $cov(I^M, I^{SOII\sim})$, which is identified from the variance formula for the sum of two random variables $I^{SOII} = I^M + I^{SOII\sim}$.

[11] Adjusted with a finite population correction, so $V(p) = (I_s^{SOII})^{-1} p(1-p)*(1- I_s^{SOII} / I^{SOII})$ where I_s^{SOII} is the SOII sample size.