AMERICAN JOURNAL OF INDUSTRIAL MEDICINE 57:1090–1099 (2014)

Capture–Recapture Estimates of the Undercount of Workplace Injuries and Illnesses: Sensitivity Analysis

Leslie I. Boden, PhD*

Background *In recent years, capture–recapture methods have been applied to state Survey of Occupational Injuries and Illnesses (SOII) and workers' compensation data, but uncertainties arise because of population differences between the two sources, likely source dependence, and concerns about the accuracy of case linkage.*
Methods *We linked SOII and workers' compensation records for California and used capture–recapture methods to estimate the proportion of injuries and illnesses involving at least 4 days away from work captured by each source. We then did a sensitivity analysis.*
Results *Assuming source independence, estimates of the proportion of injuries and illnesses involving at least 4 days away from work captured by the SOII varied from 42.4% to 49.0%, while workers' compensation estimates were between 76.9% and 77.6%. Re-estimating SOII capture rates assuming source dependence (OR = 3) reduced capture estimates substantially.*
Conclusions *Estimated capture rates remained low after changing several of the underlying assumptions. Positive source dependence had the greatest impact.* Am. J. Ind. Med. 57:1090–1099, 2014. *© 2013 Wiley Periodicals, Inc.*

KEY WORDS: occupational injury and illness reporting; capture–recapture analysis; injury surveillance; sensitivity analysis

BACKGROUND

The Survey of Occupational Injuries and Illnesses (SOII) is an annual, nationally representative survey of over 230,000 establishments conducted by the U.S. Department of Labor, Bureau of Labor Statistics (BLS). It is the most important source of detailed national estimates of the number and incidence rate of workplace injuries and illnesses in the United States. The SOII allows estimates of injuries and illnesses by industry and size of the employer; age, gender, and occupation of the worker; and nature, type, and body part of injury or illness. As such, it is a valuable resource for everybody wishing to understand the magnitude and distribution of workplace risks in the United States. The SOII also provides overall injury and illness information by state, as well as limited breakdowns by establishment, worker, and injury characteristics. Provision of more detailed information is limited by the statistical power of the survey and by confidentiality concerns.

However, concerns have been raised about the extent to which the SOII captures the injuries and illnesses it was designed to capture. Smith et al. [2005] derived injury rates for private industry from the National Health Interview Survey (NHIS) that were 1.4 times the BLS estimates. More recently, Rosenman et al. [2006] and Boden and Ozonoff [2008] have estimated wide variation in SOII reporting in several states, with most state estimates of the SOII capture rate ranging from 50% to 70%. Estimates suggest that

Boston University School of Public Health, Department of Environmental Health, Boston, Massachusetts
Contract grant sponsor: U.S. Department of Labor, Bureau of Labor Statistics.
Disclosure Statement: The authors report no conflicts of interests.
*Correspondence to: Leslie I. Boden, PhD, Boston University School of Public Health, Department of Environmental Health, 715 Albany Street, Boston, MA 02118.
E-mail: lboden@bu.edu
Accepted 2 August 2013
DOI 10.1002/ajim.22247. Published online 10 September 2013 in Wiley Online Library (wileyonlinelibrary.com).

© 2013 Wiley Periodicals, Inc.

workers' compensation capture tends to be more complete, but typically is less than 80%.

Additional concerns have been raised about whether there is substantial variation in reporting across states [Burton and Spieler, 2001; Boden and Ruser, 2003; Boden and Ozonoff, 2008; Mendeloff and Burns, 2013]. State injury rates can be compared, but a higher SOII injury rate in one state compared with another could be caused either by more hazardous conditions or by better reporting. Similar questions have been raised about reporting to state workers' compensation systems. Injury rates and costs in those systems are often used for interstate comparisons, but differential reporting and take-up of benefits can give rise to misleading inferences.

This research is based on an application of capture–recapture analysis, an important tool that uses two or more sources of data on a population (in this case the population of injured workers) to develop an estimate of the full population, including those population members observed by none of the data sources. Given certain assumptions, this estimate is unbiased. Although originally developed in wildlife biology and demography, capture–recapture methods have been used to generate improved estimates of drug use, homelessness, infectious diseases, and diabetes, to name a few [Yip et al., 1995]. In the occupational health and safety arena, they have been used to estimate occupational fatality rates [Cormack et al., 2000], the incidence of work-related musculoskeletal disorders [Morse et al., 2001], carpal tunnel syndrome [Maizlish et al., 1995], and amputations [Samant et al., 2012; Tak et al., 2013]; and the period prevalence of occupational diseases [Alho, 1990]. They also have been used to estimate overall non-fatal occupational injury and illness rates [Rosenman et al., 2006; Boden and Ozonoff, 2008].

There are significant challenges to developing good work-related injury and illness estimates from capture–recapture methods using BLS SOII and workers compensation data. This article applies sensitivity analysis to address some of these challenges. The first is identifying single-establishment and multiple-establishment firms. For purposes of this study, an establishment has a single physical location and is in a single industry. In this study, we identify a firm, which may be composed of one or more establishments, by a federal Employer Identification Number (EIN). In workers' compensation, injuries are typically reported at the firm level. However, the SOII samples primarily at the establishment level. For this reason, single- and multiple-establishment firms must be analyzed differently. To do this, we must identify which firms fall into which category. However, the data linking establishments and firms are not nearly 100% accurate [Handwerker and Mason, 2013]. This study has developed alternative estimates to approach to this issue.

Another issue addressed in this study is whether to include permanent partial disability (PPD) cases lacking evidence of meeting the SOII days-away-from work condition in the workers' compensation analysis sample. In principle, it would seem that these cases should be excluded because SOII eligibility requires at least one full day off work following the date of injury [Oleinick and Zaidman, 2010]. However, PPD cases often involve legal dispute. Unlike temporary disability cases, which typically do not involve dispute, and for which payment is begun shortly after injury, PPD cases are often settled or adjudicated long after the injury is reported. In addition, payment for temporary total disability (TTD) may be folded into PPD payments and thus never separately recorded. This is a particular concern in California, where about 40% of lost-time cases in 2007–2008 were PPD cases. Of these PPD cases, 35% (about 14% of all lost-time cases) never received any benefits categorized as temporary disability. Because of our uncertainty about whether PPD cases lacking TTD benefits actually involved one or more days away from work, we provide estimates both with and without these cases.

Basic estimates rely on the assumption of source independence, that reporting a case to the SOII is independent of reporting it to the workers' compensation system. Yet we expect positive source dependence—injuries reported to the workers' compensation system are more likely to be reported to the SOII than injuries not reported to the workers' compensation system.

There are several reasons for this. Reporting to both systems is strongly influenced by employers. In some workplaces, the same individual may be responsible for filling out the BLS survey and for monitoring workers' compensation claims [Wuellner and Bonauto, 2013]. Even if this is not the case, an employer may believe that a workplace injury that is not a compensable workers' compensation claim should not be reported to the BLS. Moreover, unless an injured worker reports an injury to a supervisor, it will very rarely enter either system. Despite these and other reasons that positive source dependence is likely, we cannot estimate the degree of dependence of the two sources without additional assumptions or a third data source [Wittes et al., 1974]. Because of this, we do a sensitivity analysis of the potential effects of source dependence, using what we consider a modest degree of source dependence, knowing what we do about the close relationship between SOII and workers' compensation reporting.

Finally, misclassification of linked cases is another possible cause of biased estimation. We examined the potential impact of misclassification of linked cases as unlinked. Such misclassification would bias downward our estimates of the completeness of reporting.

Injuries considered in scope for this analysis include workers' compensation non-fatal lost-time cases and SOII cases involving at least 4 days away from work, restricted work, or job transfer. We excluded firms that did not report as part of the SOII. We also eliminated from the study injuries

reported by temporary employment agencies, agricultural establishments with fewer than 11 employees, membership organizations, and tribal businesses. These restrictions, described in more detail below, limit the scope to injuries reportable to both sources.

METHODS

Gathering and Preparing Source Data

The source data come primarily from the California SOII for 2007 and 2008 and the California Workers' Compensation Information System (WCIS) for 2006–2008.[1] In principle, all workers' compensation insurers and self-insured employers are required to report all workers' compensation cases to the WCIS. These include both lost-time cases (those involving more than 3 days of restricted or lost work, permanent disability, or death) and medical-only cases (those involving payment for injury-related medical costs but not qualifying as lost-time cases). The SOII collects data on non-fatal injuries from a stratified random sample of establishments. Reportable cases involve at least one full day away from work after the date of injury. Reports are due shortly after the report year.

Aside from checking ranges and otherwise cleaning the data, new data elements were created to facilitate automated comparisons of cases in the two data sources. All lower-case letters in names and addresses were transformed to uppercase. Then, computer programs modified from those used by the United States. Bureau of the Census were used to standardize first names and addresses. Also, Soundex last names were created, as were shortened versions of the employer's name and address. Soundex produces a phonetic version of the last name so that misspellings will have less of an impact on the third step, linking cases from the two data sources.

Linking workers' compensation and SOII cases

The purpose of this step is to determine which injuries were reported to both systems, which only to workers' compensation, and which only to the SOII. We linked using the following characteristics: (1) either date of injury or month of injury within injury year, (2) first name, (3) last name or Soundex value of last name, (4) EIN, (5) date of birth or age, (6) employer name, (7) the employer address or the simplified address (number at the beginning plus the first full word), (8) employer zip code or city, and (9) worker's sex. Industry was not available in the workers' compensation data, so it was not used to link. Nature of injury and part of body injured were available in both datasets, but we did not use them for two reasons. First, they were coded using different coding systems, and second, past experience led us to believe that, given these differences, there was also much inconsistent and inaccurate coding of these data elements within each system. Linkpro 3.0 software (© InfoSoft, Inc., Winnipeg, Manitoba, Canada) was used for the linkage. Initially, the linkage was deterministic, requiring nine of the following to match: date of birth, age, month of injury, date of injury, first name, first initial, last name, Soundex of last name, EIN, employer name, employer address, employer city. After removing the deterministically linked cases, a probabilistic linkage was done [Fellegi and Sunter, 1969; Copas and Hilton, 1990]. The probabilistic linkage method produces weights that rank pairs according to the probability that potential links are actually the same injury. The potential linkages were ranked by weight. By inspection, we determined an upper range of weights that were definitely linked and a lower range that were not. The remaining potentially linked cases were examined clerically by two people who independently categorized pairs as linked or unlinked. Cases for which both reviewers agreed (Kappa $= 0.84$) were linked were added to the other linked cases. The linkage status of cases for which the coders did not agree was determined by a tie-breaking rule. After this, we removed the linked cases from both datasets and used Linkpro 3.0 on the new dataset to look for additional linked cases. Of cases determined to be linked, 74.4% were deterministically linked, and the rest were probabilistically linked.

The major concern about linkage misclassification is that linked cases would be classified as unlinked and thus would bias downward the estimated probability that cases were reported. This was a particular concern among cases for which the two clerical reviewers did not agree. To see if our results were likely to be sensitive to this potential source of misclassification, we reclassified all these cases as linked.

Harmonizing Dataset Scope

The workers' compensation data are a census of all cases, while the SOII is based on a sample of establishments. Capture–recapture requires that the data be drawn from the same population, so the workers' compensation data should be restricted to the establishments in the SOII. Generally speaking, employers do not report to the workers' compensation system by establishment, but by firm, which can be identified by the federal Employer Identification Number (EIN). An initial step in harmonizing the data from both sources is to remove firms without SOII-reporting establishments from the workers' compensation sample. We first determined which firms in the workers' compensation sample to keep on the basis of linked cases. If the same injury was reported to both workers' compensation and the SOII, we assumed that the injury occurred in the workers'

[1] 2006 workers' compensation cases were only used to identify cases with dates of injury in 2006 but for which lost time did not begin until after 2006.

compensation firm. A substantial number of SOII establishments reported no injuries, so this method could not be used for them. For these establishments, we included workers' compensation firms in the research dataset if the two sources agreed on EIN, the first six characters of the employer name, the first six characters of the employer's address, and the employer city.

For many cases, either no EIN or an invalid EIN (like 999999999 or 123456789) was provided to workers' compensation. We attempted to replace these by finding another claim with the same first 12 characters of the employer name and a valid EIN. For those cases, we replaced the missing or invalid EIN with the valid EIN. (In our experience, such a large number of missing employer identifiers in the state workers' compensation database is unique to California.)

The SOII includes only cases where workers do not return to work for at least one full day after the day of the injury. In contrast, the California workers' compensation database includes both lost-time and medical-only cases. Lost-time cases involve more than 3 days of total or partial disability. Medical-only cases involve 3 or fewer days, but the number of days is not specified. To harmonize the data from both sources, we discarded SOII cases involving fewer than 4 days for which the worker was away from work, had a work restriction, or had a job transfer because of injury unless they were linked to a lost-time workers' compensation case. We also discarded workers' compensation medical-only cases and those with only temporary partial disability.

In one analysis, we kept only workers' compensation cases with either TTD or permanent total disability (PTD) payments, discarding cases with only temporary partial disability or PPD payments unless they were linked to SOII cases. In another analysis, we kept unlinked PPD cases even if no TTD benefits were paid.

Coverage of some establishments differs between the two sources. Temporary employment agencies were eliminated because injuries to their employees are reported differently to workers' compensation (by the temporary employment agency) and to SOII (by the establishment at which they worked). Agricultural establishments with fewer than 11 employees are not surveyed for the SOII national estimates. Finally, membership organizations and tribal businesses (primarily casinos) have spotty coverage by workers' compensation programs. We exclude these industries from our analysis, reducing coverage in favor of analyzing industries with relatively uniform reporting requirements.

Injuries in some establishments were reported to the workers' compensation system under an insurer name, rather than an employer name. These cases and their employers were eliminated because cases in the SOII sampling frame and those not in the sampling frame could not be distinguished.

To reduce the reporting burden, firms with a large number of expected injuries are asked to report to the SOII only a subsample of those injuries—only injuries occurring in specific months or on specific days of the month. To harmonize the workers' compensation and SOII samples, unlinked workers' compensation injuries not conforming to the subsampling rule were removed from the sample.

Distinguishing single-establishment from multiple-establishment firms

The remaining major challenge for capture–recapture analysis was that workers' compensation cases are reported by firm, while the SOII is a sample of establishments. As a consequence, we could identify firms in the workers' compensation data with establishments that reported to the SOII, but we could not identify reporting and non-reporting establishments within those firms. For multiple-establishment firms, an injury may have been unrecorded in the SOII because an establishment failed to report or because the establishment was not included in the SOII subsample. This means that, for multiple-establishment firms, unlinked workers' compensation cases may have been in unsampled establishments and should not have been considered as missed by the SOII.

We approached this problem in two steps. First, we distinguished single- and multiple-establishment firms. With the help of BLS economists, we used the SOII EIN linked to the BLS Longitudinal Database (LDB) [Pivetz et al., 2001] to distinguish between single- and multi-establishment employers. In addition, because we also had EIN information from workers' compensation, and this information wasn't always the same as that in the SOII, we also developed a single-establishment indicator from the workers' compensation database. (For a detailed discussion of the difficulty of linking establishments to firms, see Handwerker and Mason [2013].) We then developed a joint single-establishment indicator using both the SOII and the workers' compensation identifiers. If either indicated that an employer has more than one establishment, the employer was identified as multiple-establishment. Otherwise, if either or both indicated that the employer has only one establishment, it was identified as single-establishment.

Using these methods, some establishments could not be identified as either single- or multiple-establishment. These establishments were dropped from the analysis. If reporting in these establishments differed from reporting in the other establishments, this could have biased our results.

Approach for single-establishment firms

The single-establishment firm analysis was relatively straightforward. Weighted values of unlinked SOII cases, unlinked workers' compensation cases, and linked cases were

calculated to account for SOII sampling and subsampling, and the Lincoln–Peterson estimator was used to calculate the estimated total number of cases under the assumption of source independence. This was done for two alternative workers' compensation samples, one that included only cases involving TTD or PTD payments, and a second that also included PPD cases in which no TTD payment was indicated.

We also used two ways of distinguishing single- from multiple-establishment firms. One used only the SOII employer single-establishment identifier. The second used the joint SOII-workers' compensation identifier.

Approach for multiple-establishment firms

Some unlinked workers' compensation cases did not have a corresponding SOII case because they were in firms sampled by the SOII but not in the specific sampled establishments. To address this, we used quarterly employment information from the LDB to derive for each firm the proportion of employment in establishments reporting to the SOII. We assumed homogeneity of injury rates and reporting rates across all establishments within a firm. (This is not likely to be generally true, but it seems reasonable as an average expectation.) Under this assumption, we imputed the number of workers' compensation injuries reported by establishments that were not in the BLS sample and adjusted the sampling weights of the unmatched workers' compensation cases to account for the expected number of unlinked workers' compensation injuries reported by unsampled establishments in multiple-establishment firms.

Several multiple-establishment firms had more than one establishment reporting to the SOII in a given year. For these establishments, we applied the same method to derive a weight adjustment, treating them as a single-establishment. However, there was another issue: it was not uncommon for sampled establishments within a firm to have different sampling weights. Because we did not know in which establishment the unlinked workers' compensation injuries occurred, the sampling weights for those injuries were indeterminate. We addressed this by devising a sampling weight for the unlinked cases that was the average sampling weight for the reported SOII cases for all sampled establishments in the firm.

Capture–Recapture Estimation

We estimated the total number of injuries, including those reported to neither source, using the Lincoln–Peterson estimator [Hook and Regal, 1995], assuming source independence. We derived estimates separately for single-establishment and multiple-establishment firms. Each of these estimates was derived using two methods of identifying firm type. These four estimates were then derived including PPD cases with no TTD payments and excluding them—eight estimates in all. We then summed the single-establishment and multiple-establishment estimates to derive four separate estimates of total injuries and capture rates. We then recalculated these estimates assuming positive source dependence. None of these estimates were covariate-adjusted. We then derived covariate-adjusted estimates using logistic regression, with worker age and sex, employer size and industry, and nature of injury and part of body and as covariates.

This study was approved by the Boston University Medical Center Institutional Review Board (IRB) for protection of human subjects. The IRB approved a waiver of informed consent. Approval was based on strict procedures to protect identifiable information, minimal disclosure risk to subjects, and impracticality of contacting individuals in the administrative data.

RESULTS

Reporting to the SOII and to Workers' Compensation, Source Independence

Table I shows capture–recapture estimates assuming source independence using two ways of harmonizing the SOII and workers' compensation data. In one case, we included all PPD cases, and in the other we excluded any PPD case that was not linked to a SOII case and lacked TTD payments. We also used two ways of distinguishing single- from multiple-establishment firms. One used only the SOII employer identifier. The second used both the SOII and the workers' compensation identifiers. If either indicated that an employer has more than one establishment, the employer was identified as multiple-establishment. Otherwise, if either or both indicated that the employer had only one establishment, it was identified as single-establishment.

From Table I, we can see that estimates of single-establishment SOII capture rates vary from 40.0% to 51.6%. For multiple-establishment firms, these estimates vary from 41.6% to 50.6%. The third panel in Table I provides an estimate of overall capture rates, which vary from 42.4% to 49.0%. The joint single-establishment indicator moves workplaces from the single to the multiple-establishment category, increases the estimated SOII reporting fraction for workplaces identified as single-establishment, and reduces this number for those identified as part of multiple-establishment firms. However, the overall estimate of the number of cases and the SOII reporting fraction changes very little. The 95% confidence intervals show that the impact of model and data uncertainties is much greater than statistical variability.

TABLE I. Capture–Recapture Estimates of Injury Reporting to the SOII Assuming Source Independence, California, 2007–2008

	SOII indicator		Joint indicator[a]	
	Percent of cases	Estimated total cases	Percent of cases	Estimated total cases
Single-establishment firms				
Including PPD-only cases[b]	40.0	196,591 (N = 23,520)	46.8	162,469 (N = 19,086)
Excluding PPD-only cases	44.1	178,246 (N = 22,838)	50.6	150,087 (N = 18,715)
Multiple-establishment firms				
Including PPD-only cases[b]	46.1	128,311 (N = 52,581)	41.6	161,827 (N = 57,329)
Excluding PPD-only cases	51.6	114,720 (N = 47,510)	47.4	142,270 (N = 51,903)
All firms				
Including PPD-only cases[b]	42.4	324,902 (N = 76,101)	44.2	324,296 (N = 76,415)
Excluding PPD-only cases	47.0	292,966 (N = 70,348)	49.0	292,356 (N = 70,618)

Note: Temporary employment agencies, membership organizations, tribal businesses, and insurers have been eliminated from this sample. The analysis sample includes SOII and workers' compensation cases with lost days greater than the workers' compensation waiting period. All eight capture–recapture tables can be reconstructed from this table and Table II. The 95% confidence interval for the estimate of total cases is between 1.4% and 1.5% for single-establishment firms, between 2.4% and 2.6% for multiple-establishment firms, and 1.3% for all firms.
[a]Determination of multiple-establishment status was based on both the EINs of SOII establishments and the EINs reported to the WC system.
[b]PPD-only cases lack TTD benefits.

Table II is parallel to Table I, but it shows the estimated workers' compensation capture rates. These estimates are insensitive to the changes in assumptions described above.

Sensitivity to Linkage Misclassification

We initially used a tie-breaking rule to determine the linkage status of cases for which coders disagreed. To determine sensitivity to the tie-breaking rule, we reclassified all these cases as unlinked. This produced increases in estimates of reporting rates of less than 0.3 percentage points for both SOII and workers' compensation. We also simulated increasing the number of linked cases by 10% and reducing SOII and workers' compensation unlinked cases by the same number. (The choice of 10% is arbitrary but high, given the care with which linkage was assessed.) This increased SOII reporting rates by 0.7–1.7 percentage points and workers' compensation reporting rates by 0.8–3.0 percentage points.

TABLE II. Capture–Recapture Estimates of Injury Reporting to Workers' Compensation Assuming Source Independence, California, 2007–2008

	SOII indicator		Joint indicator[a]	
	Percent of cases	Estimated total cases	Percent of cases	Estimated total cases
Single-establishment firms				
Including PPD-only cases[b]	74.1	196,591 (N = 23,520)	73.5	162,469 (N = 19,086)
Excluding PPD-only cases	74.1	178,246 (N = 22,838)	73.5	150,087 (N = 18,715)
Multiple-establishment firms				
Including PPD-only cases[b]	81.3	128,311 (N = 52,581)	81.8	161,827 (N = 57,329)
Excluding PPD-only cases	81.3	114,720 (N = 47,510)	81.8	142,270 (N = 51,903)
All firms				
Including PPD-only cases[b]	77.0	324,902 (N = 76,101)	77.6	324,296 (N = 76,415)
Excluding PPD-only cases	76.9	292,966 (N = 70,348)	77.5	292,356 (N = 70,618)

Note: Temporary employment agencies, membership organizations, tribal businesses, and insurers have been eliminated from this sample. The analysis sample includes SOII and workers' compensation cases with lost days greater than the workers' compensation waiting period. All eight capture–recapture tables can be reconstructed from this table and Table I. The 95% confidence interval for the estimate of total cases is between 1.4% and 1.5% for single-establishment firms, between 2.4% and 2.6% for multiple-establishment firms, and 1.3% for all firms. Ns are the same as in Table I.
[a]Determination of multiple-establishment status was based on both the EINs of SOII establishments and the EINs reported to the WC system.
[b]PPD-only cases lack TTD benefits.

The highest SOII reporting estimate for all firms rose to 49.8%, and the highest workers' compensation rate rose to 78.7%.

Source Dependence

It is highly likely that reporting to the SOII and to the WCIS is not independent. However, using two sources, we cannot estimate the degree of source dependence. Thus, if our estimates of the counts of linked and unlinked injuries are correct, the results presented so far are likely to substantially overestimate the proportion of injuries captured by both sources. To demonstrate how source dependence might affect the results of a capture–recapture analysis, we present the results of the estimates in Tables I and II, but now under the assumption of positive source dependence, with an odds ratio of three for reporting to the SOII if a case has been reported to the WCIS cases not been reported to the WCIS. These results are reported in Tables III and IV. They show that relaxing the assumption of source dependence can reduce the estimated SOII capture percentages for all firms about 9 percentage points and workers' compensation capture rates about 14–16 percentage points in each case. Another way of stating this is that capture rates for all firms declined about 20% for both SOII and workers' compensation reporting. We also calculated estimated capture rates for an odds ratio of five (tables not shown). This drove them even lower, with the SOII estimated to capture about 1/3 and workers' compensation about $^1/_2$ of all cases. An odds ratio of two (tables not shown) reduced estimated SOII capture rates by about 5 percentage points and workers' compensation capture rates by 8 to 9 percentage points.

The Impact of Capture Heterogeneity

Heterogeneity of capture rates by individual, injury, or establishment characteristics can affect estimated capture rates, so we also calculated an adjusted overall capture rate, allowing linkage probabilities to vary by worker sex, age, and tenure, the part of body injured and the nature of injury (sprain, strain, fracture, etc.), whether the injury occurred in the last 45 days of the year, whether the worker lost at least 14 days of work, the size and industry of the employer, and whether the employer was private or public. To do this, we estimated multinomial logistic regressions with covariates available in both the SOII and workers' compensation data. Separate estimates were made for single- and multiple-establishment firms, using the CATMOD procedure in SAS© 9.3. We found differences between the crude capture rates presented in Tables I through IV and the heterogeneity-adjusted capture rates, but they were all less than 2 percentage points. For example, using the joint SOII-WCIS single-establishment indicator and excluding PPD cases lacking TTD, the SOII capture rate for single-establishment firms in Table I is 50.6%. The adjusted estimate is 51.2%. For the parallel multiple-establishment firms, the SOII estimated capture rate in Table I is 53.9%, and the adjusted estimate is 55.8%.

DISCUSSION

We used two methods to assign firms as single- or multiple-establishment. These alternatives made virtually no difference in the estimated capture rate of the workers' compensation system. They changed the estimated SOII

TABLE III. Capture–Recapture Estimates of Injury Reporting to the SOII Source Dependence, Odds Ratio = 3, California, 2007–2008

	SOII indicator		Joint indicator[a]	
	Percent of cases	Estimated total cases	Percent of cases	Estimated total cases
Single-establishment firms				
Including PPD-only cases[b]	30.5	257,641	36.5	208,273
Excluding PPD-only cases	34.2	229,809	40.1	189,331
Multiple-establishment firms				
Including PPD-only cases[b]	38.4	154,189	34.3	196,242
Excluding PPD-only cases	43.7	135,511	39.7	169,559
All firms				
Including PPD-only cases[b]	33.4	411,830	35.4	404,514
Excluding PPD-only cases	37.7	365,321	40.0	358,891

Note: Temporary employment agencies, membership organizations, tribal businesses, and insurers have been eliminated from this sample. The analysis sample includes SOII and workers' compensation cases with lost days greater than the workers' compensation waiting period. All eight capture–recapture tables can be reconstructed from this table and Table IV. The 95% confidence interval for the estimate of total cases is about 1.1% for single-establishment firms, between 2.1% and 2.3% for multiple-establishment firms, and 1.0 to 1.1% for all firms. Ns are the same as in Table I. The unweighted Ns (number of observations) are the same as in Tables I and II.

[a]Determination of multiple-establishment status was based on both the EINs of SOII establishments and the EINs reported to the WC system.

[b]PPD-only cases lack TTD benefits.

TABLE IV. Capture–Recapture Estimates of Injury Reporting to Workers' Compensation Source Dependence, Odds Ratio = 3, California, 2007–2008

	SOII indicator		Joint indicator[a]	
	Percent of cases	Estimated total cases	Percent of cases	Estimated total cases
Single-establishment firms				
Including PPD-only cases[b]	56.6	257,641	57.3	208,273
Excluding PPD-only cases	57.5	229,809	58.3	189,331
Multiple-establishment firms				
Including PPD-only cases[b]	67.6	154,189	67.4	196,242
Excluding PPD-only cases	68.8	135,511	68.6	169,559
All firms				
Including PPD-only cases[b]	60.7	411,830	62.2	404,514
Excluding PPD-only cases	61.7	365,321	63.2	358,891

Note: Temporary employment agencies, membership organizations, tribal businesses, and insurers have been eliminated from this sample. The analysis sample includes SOII and workers' compensation cases with lost days greater than the workers' compensation waiting period. All eight capture–recapture tables can be reconstructed from this table and Table III. The 95% confidence interval for the estimate of total cases is about 1.1% for single-establishment firms, between 2.1% and 2.3% for multiple-establishment firms, and 1.0 to 1.1% for all firms. The unweighted Ns (number of observations) are the same as in Tables I and II.

[a]Determination of multiple-establishment status was based on both the EINs of SOII establishments and the EINs reported to the WC system.

[b]PPD-only cases lack TTD benefits.

capture rates by 6–7 percentage points for single-establishment firms and 4–5 percentage points for multiple-establishment firms. However, the overall estimated SOII capture rate only changed by about 2 percentage points.

From inspecting cases for which they differ, we believe that the joint indicator is better at distinguishing single- from multiple-establishment firms. However, we know from observing the data and from another study [Wuellner and Bonauto, 2013] that even this more conservative method is likely to classify some multi-establishment firms as single-establishment. We have seen, among others, a major commercial airline and a retail chain classified as single-establishment. We carefully examined the data, so we were able to reduce this misclassification, but we doubt that we eliminated it. More study would be needed to determine how much this could affect estimates of the SOII undercount.

Excluding PPD-only cases had a larger impact—about 5 percentage points averaged over all firms. Still, estimates of SOII capture rates in California remained below 50%. Oleinick and Zaidman [2010] expressed concern that including PPD cases that lacked TTD payments would lead to an underestimate of the SOII reporting rate. This is because the SOII requests reporting for cases involving at least one day away from work. Table V suggests that this is a reasonable concern. Clearly, although many PPD cases lacking TTD payments are reported to the SOII, they are considerably less often reported than are PPD cases with TTD. The disparity in reporting is great, providing support for

TABLE V. Lack of TTD Payments and Estimated Number and Proportion of PPD Injuries Reported to the SOII: Joint SOII-WCIS Single-Establishment Indicator

	1+ TTD days	No TTD days	Total
Single-establishment firms (N = 6,240)			
Not reported to SOII	14,951 (52.1%)	9,103 (84.8%)	24,054 (61.0%)
Reported to SOII	13,747 (47.9%)	1,631 (15.2%)	15,378 (39.0%)
Total	28,698 (100.0%)	10,734 (100.0%)	39,432 (100.0%)
Multiple-establishment firms (N = 20,749)			
Not reported to SOII	19,081 (57.3%)	10,992 (90.3%)	30,073 (66.1%)
Reported to SOII	14,194 (42.7%)	1,178 (9.7%)	15,372 (33.8%)
Total	33,275 (100.0%)	12,170 (100.0%)	45,445 (100.0%)

Note: Determination of single-establishment status was based on the EINs of both SOII and WCIS establishments. Temporary employment agencies, membership organizations, tribal businesses, and insurers have been eliminated from this sample. The analysis sample includes SOII and workers' compensation cases with lost days greater than the workers' compensation waiting period.

the hypothesis that many of these cases are not reportable under the SOII rules. Still, excluding these cases from the analysis does not qualitatively affect conclusions about the SOII undercount.

Our results do not appear to be sensitive to linked cases that were misclassified. All probabilistically linked cases were clerically reviewed by two people. They disagreed about the linkage status of only a small number of cases that were included in the analysis datasets. Treating these cases as linked had a very small impact on final capture–recapture estimates.

A substantial amount of underreporting may derive from workers' not reporting injuries to their employers. There are many possible reasons for this, including fear of job loss, pressure by employers or fellow workers not to report, lack of knowledge about workers' compensation, discomfort with entering the workers' compensation system, and workers' compensation coverage restrictions. Injuries that are unknown to the employer are not detected in this analysis. In fact, they are a cause of positive source dependence, which implies that capture–recapture analysis assuming source independence overestimates the proportion of injuries reported.

In addition, employers may use workers' compensation records as a source of information for filling out the OSHA log. Even if they do not do so, they may think that they are not required to report to the SOII injuries that they believe do not qualify for workers' compensation benefits. Both would lead to positive source dependence.

Two-source reporting cannot detect source dependence [Wittes et al., 1974]. Nor can employer interviews determine whether workers are reporting their injuries. Yet source dependence appears to be a very important potential source of uncertainty in estimating the level of workplace injury reporting. This suggests that worker interviews or population-based worker surveys (possibly adding questions to existing population surveys) may be excellent supplements to studies using administrative data.

Limitations

We developed a method of dealing with the possibility that many injuries reported to the workers' compensation system from multiple-establishment employers may be outside the scope of the SOII establishment sample. This method down-weighted these cases, but it relies on strong assumptions, including assumptions about the accuracy of the LDB employment data. Moreover, the fact that many injuries are in multi-establishment employers for which establishments do not have a unique sampling weight creates additional model uncertainty in the estimates derived in this report. In addition, the calculated confidence intervals only take into account sampling error, not the error induced by the derived weights for multiple-establishment firms.

Attempts to address the problem of missing and invalid EINs in the workers' compensation data left a substantial number of EINs as unknown, and may have misclassified some EINs. Claims that should have been in the analysis sample may have been left out, and others that should have been excluded may have been kept.

Also, we know from other studies that reporting of chronic occupational illnesses is very low and unreliable in both sources we used. Such underreporting is likely to be very low in both sources and highly correlated between sources. As a consequence, two-source capture–recapture analysis is likely to provide a very poor estimate of these conditions.

Finally, the two sources may report different numbers of days off work for the same injury in the two sources. We have kept linked cases with different reported lost time, but have dropped unlinked SOII cases when they did not appear to have more than 3 days off work. Because of differences in reporting between the SOII and workers' compensation, we may have dropped cases that should have been retained and kept some that should have been dropped. We only dropped unlinked cases with 3 or fewer days off work from the analysis, which biased our estimates toward higher capture rates. Still, if unlinked cases with 3 or fewer days off work were reported as having more than 3 days off, this would bias estimated capture rates downward.

CONCLUSIONS

This article has explored the potential impact of several sources of uncertainty about capture–recapture estimates of reporting of non-fatal workplace injuries and illnesses. We have shown that two methods of distinguishing single- and multiple-establishment firms had little impact on estimated SOII reporting rates—at most 2 percentage points over all firms. Excluding PPD-only cases had a greater impact, about 5 percentage points. Effects were much smaller for workers' compensation reporting. Taking into account both these sources of uncertainty, overall estimates of SOII reporting varied from 42.4% to 49%, and workers' compensation estimates varied from 76.9% to 77.6%. Qualitatively, using different approaches, conclusions remained similar: the SOII missed about 40–50% of claims, and the workers' compensation system missed about 20–25%. A moderate amount of linkage misclassification also would not have qualitatively changed our conclusions. Results were more strongly affected by assumptions about the strength of positive source dependence. Assuming an odds ratio of 3 instead of source independence had about twice the impact on SOII capture rates as excluding PPD cases and an even more substantial relative impact on workers' compensation capture rates. Notably, allowing for source independence drives estimated capture rates even farther away from 100%.

The results of capture–recapture analysis for the state of California were only modestly affected by attempts to model several sources of uncertainty. These results continue to imply that current estimates of the extent of non-fatal occupational injuries and illnesses substantially understate their magnitude. They support efforts to understand the sources of the undercount and to improve our understanding of the true magnitude of this important public health issue.

ACKNOWLEDGMENTS

We thank the Bureau of Labor Statistics and the State of California Department of Industrial Relations for their assistance in providing data and insights into their systems. Opinions presented here do not reflect those of either organization. Particularly helpful were Christine Baker, Martha Jones, Lachlan Taylor, John Ruser, Brooks Pierce, Nichole Nestoriak, and Eric Sygnatur. Al Ozonoff, Letitia Davis, and David Bonauto provided helpful suggestions on drafts of this article. Matt Frederick engaged in helpful discussions on data issues, and Steven Evans provided excellent statistical programming.

REFERENCES

Alho JM. 1990. Logistic-regression in capture recapture models. Biometrics 46:623–635.

Boden LI, Ozonoff A. 2008. Capture-recapture estimates of nonfatal workplace injuries and illnesses. Ann Epidemiol 18:500–506.

Boden LI, Ruser JW. 2003. Workers' compensation "reforms," choice of medical care provider, and reported workplace injuries. Rev Econ Stat 85:923–929.

Burton JF, Spieler EA. 2001. Workers' compensation and older workers. Ensuring Health and Income Security for an Aging Workforce, 41–83.

Copas JB, Hilton FJ. 1990. Record linkage—Statistical-models for matching computer records. J R Stat Soc Ser A Stat Soc 153:287–320.

Cormack RM, Chang YF, Smith GS. 2000. Estimating deaths from industrial injury by capture-recapture: A cautionary tale. Int J Epidemiol 29:1053–1059.

Fellegi IP, Sunter AB. 1969. A theory for record linkage. J Am Stat Assoc 64:1183.

Handwerker EW, Mason LG. 2013. Linking firms with establishments in BLS microdata. Mon Labor Rev 136:14–22.

Hook EB, Regal RR. 1995. Capture-recapture methods in epidemiology: Methods and limitations. Epidemiol Rev 17:243–2264.

Maizlish N, Rudolph L, Dervin K, Sankaranarayan M. 1995. Surveillance and prevention of work-related carpal tunnel syndrome: An application of the Sentinel Events Notification System for Occupational Risks. Am J Ind Med 27:715–729.

Mendeloff J, Burns R. 2013. States with low non-fatal injury rates have high fatality rates and vice-versa. Am J Ind Med 56:509–519.

Morse T, Dillon C, Warren N, Hall C, Hovey D. 2001. Capture-recapture estimation of unreported work-related musculoskeletal disorders in Connecticut. Am J Ind Med 39:636–642.

Oleinick A, Zaidman B. 2010. The law and incomplete database information as confounders in epidemiologic research on occupational injuries and illnesses. Am J Ind Med 53:23–36.

Pivetz T, Searson M, Spletzer J. 2001. Measuring job and establishment flows with BLS longitudinal microdata. Mon Labor Rev 124:13–20.

Rosenman KD, Kalush A, Reilly MJ, Gardiner JC, Reeves M, Luo ZW. 2006. How much work-related injury and illness is missed by the current national surveillance system? J Occup Environ Med 48:357–365.

Samant Y, Parker D, Wergeland E, Lund J, Westin S. 2012. Estimating work-related amputations in the Norwegian manufacturing sector: a 10-year retrospective study based on two-source capture-recapture method. Int J Occup Environ Health 18:292–298.

Smith GS, Wellman HM, Sorock GS, Warner M, Courtney TK, Pransky GS, Fingerhut LA. 2005. Injuries at work in the US adult population: Contributions to the total injury burden. Am J Public Health 95:1213–1219.

Tak S, Grattan K, Boden LI, Ozonoff A, Davis L. 2013. Impact of differential injury reporting on the estimation of the total number of work-related amputations. Am J Ind Med XX.

Wittes JT, Colton T, Sidel VW. 1974. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. J Chronic Dis 27:25.

Wuellner S, Bonauto D. 2013. Why injury reports in the BLS Survey of Occupational Injuries and Illnesses differ from WC claims data: Interviews with survey respondents. Am J Ind Med XX.

Yip P, Bruno G, Tajima N, Seber G, Buckland S, Cormack R, Unwin N, Chang Y, Fienberg S, Junker B, Laporte R, Libman I, McCarty D. 1995. Capture-recapture and multiple-record systems estimation II: Applications in human diseases. Am J Epidemiol 142:1059–1068.