Using Capture-Recapture Analysis to Identify Factors Associated with Differential Reporting of Workplace Injuries and Illnesses

Leslie I. Boden[1], Nicole Nestoriak[2], Brooks Pierce[2]
[1]Boston University School of Public Health, 715 Albany Street, Boston, MA 02118
[2]US Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington DC 20212

Abstract
Recent work comparing the Bureau of Labor Statistics' Survey of Occupational Injuries and Illnesses (SOII) to Workers' Compensation (WC) claims databases concludes that the SOII substantially undercounts cases. We use linked WC-SOII data from Wisconsin to describe which cases are more likely to be accurately captured in the SOII. The SOII capture rate is higher for relatively acute injuries such as fractures, and is lower for injuries that are less readily identifiable as work-related such as inflammation or carpal tunnel, or for injuries in which WC claims were made substantially after injury incidence or after the year of injury. These findings further our understanding of the strengths and weaknesses in both the methodology used to measure the undercount and in the current data collection methods used by the SOII.

Key Words: workplace injuries and illnesses, capture-recapture

1. Introduction

There is substantial evidence that many workplace injuries go unreported and uncounted in the U.S. This is a concern because the accuracy of workplace injury and illness reporting is important for prevention efforts, as efforts to improve workplace safety rely on our understanding of the risks faced by workers.[1]

The view that injuries are undercounted is based in part on recent studies comparing injuries reported to different surveillance sources. The main data sources for these studies include the U.S. Bureau of Labor Statistics' Survey of Occupational Injuries and Illnesses (SOII), workers' compensation (WC) administrative data, and hospitalization or physician reports. For example, Rosenman et al. (2006) compare case reports from several sources for Michigan over the 1997-2001 period. That study determines that the SOII, which is the most prominent and widely cited source for workplace safety statistics, counts as few as one-third of injuries and illnesses. Similarly, Boden and Ozonoff (2008) compare SOII and WC data for several states over roughly the same time frame and conclude that the SOII detects between 50 percent and 75 percent of cases in the states studied.

1 See for example congressional testimony on under-reporting of workplace injuries and illnesses at http://edlabor.house.gov/hearings/2008/06/hidden-tragedy-underreporting.shtml. Ruser (2008) provides an overview of reporting issues.

This study follows on Boden and Ozonoff (2008) by documenting for one state, Wisconsin, the characteristics of cases that are more likely to be captured, or more likely to be missed, by the SOII or by the WC system. Rather than focusing on the overall magnitude of the undercount, this study attempts to identify the factors that have the greatest impact on capture propensities. As emphasized in Azaroff (2002), a wide variety of incentives and informational barriers can affect reporting propensities to any particular surveillance source. These factors will naturally be reflected in the types of cases that are especially easy or especially difficult to report and count. Identifying situations where the undercount is estimated to be larger may help us understand reporting incentives or identify possible areas for improvement in the surveillance data sources.

2. Data Sources

The SOII is an annual establishment survey of workplace injuries and illnesses. The survey scope includes private sector and state and local government employers. Farms with fewer than 11 employees and private households are excluded. Because the SOII surveys employers, unincorporated self-employed workers are out of scope. Data for certain mining and railroad activities are not collected via the survey, but are reported to BLS by regulatory agencies charged with that task. SOII samples are drawn several months prior to the survey year, and sampled establishments are notified that they must report on new injury and illnesses occurring over the course of the survey calendar year. An employer with multiple establishments within a state may have some, all, or none of its separate establishments sampled. Respondents report information from on-site injury logs maintained as part of the Occupational Safety and Health Administration's (OSHA) record-keeping requirements.[2] Data for a given survey year are reported to BLS in the first half of the following year.

For more serious injury or illness cases—those involving at least 1 day away from work beyond the date of injury or onset of illness—the SOII collects more detailed information describing the incident and the affected employee. The SOII program refers to these cases as "days away from work" cases. Collected information includes the nature and source of the injury or illness, the part of the body affected, and the date of injury or onset of illness, as well as the employee's name, date of birth, and gender. These fields, as well as information on the employer, are used to help identify cases for the purposes of matching SOII records to records from other surveillance sources. For most establishments, the SOII intent is to record a census of days away from work cases. For those establishments with a large expected or realized number of cases, the SOII subsamples cases within the establishment by date of injury or illness onset. This case subsampling reduces respondent burden.

In Wisconsin in 2000, SOII reports a days away from work case rate of 2.5 per 100 full-time equivalent workers.[3] Most SOII respondents report zero days away from work cases. Among respondents reporting some days away from work cases there is substantial skewness, with many reporting only one or two cases and some reporting dozens or even hundreds of cases. Injuries tend to be much more commonly reported than illnesses.

[2] Respondents exempt from OSHA record-keeping were "pre-notified", meaning the SOII sent them OSHA record-keeping forms prior to the survey year.

[3] This corresponds to approximately 49,500 private sector injury and illness cases requiring days away from work, see http://www.bls.gov/iif/oshwc/osh/os/pr006wi.pdf and http://www.bls.gov/iif/oshwc/osh/os/pr007wi.pdf, (accessed July 12, 2010).

In contrast to the SOII, WC administrative systems are unique to each state. Wisconsin's WC system mandates coverage of nearly all private-sector workers. WC typically covers almost all medical expenses arising from a work-related injury or illness, covers portions of lost earnings due to temporary injuries or illnesses if the duration of the injury or illness exceeds a minimum waiting period, and provides partial or total disability payments in the event of permanent disability. Injury and illness cases in Wisconsin from 1998 to 2001 received benefits for lost earnings under WC if the disability lasted 4 or more days. Employees have two years to report a workplace injury to their employers, although most injuries are reported much earlier. Some traumatic injuries (vision loss, total loss of a hand or arm, permanent brain injury, etc.) and some occupational diseases (carpal tunnel syndrome, hearing loss, etc.) have no time limit for filing a claim.

Under the WC system, benefits may be requested by workers but disputed by the employer.[4] An employer may believe a given injury is not work related, or the employer may dispute the degree or duration of disability. In such cases the employee may request that the State office of WC resolve the dispute via a hearing before an administrative law judge. Negotiated settlements are possible. The WC data used here include contested cases and negotiated settlements. These are identified separately from other cases when they result in lump-sum settlements or are awarded by an administrative law judge. Otherwise, it may not be possible to identify them.

To give a sense for magnitudes, the Wisconsin WC system reported on average about 50,000 lost-time claims per year over the 2000–2006 period. Of these, about 18 percent (an annual average of about 9,200 claims) were marked as denials, or as injuries or illnesses that did not require days away from work, or as non-compensable cases. About 13.6 percent (6,800) of claims were litigated annually.[5]

Many data elements in the WC administrative data are broadly similar to those collected in the SOII. WC data include information on the worker including name and date of birth; information on the injury or illness, such as date of injury and a coding of the nature of the injury or illness and part of body affected; and, information on the employer such as name, address, and identifiers used in state and federal reporting systems. This similarity of data fields aids in linking and comparing the data, described below.

However, the two systems have different purposes and are constructed differently. The WC system is an ongoing system designed to track case dispositions and benefit payments, whereas the SOII is designed to be a scientifically valid sample of new cases occurring in a given calendar year. The different purposes can make the comparison exercise difficult in some ways. For example, the SOII is constructed around the concept of recording cases at a particular worksite whereas the WC system records employer information but has little reason to note worksite location, and this makes it difficult to ascertain for employers with multiple worksites whether a WC case occurs at a location sampled by the SOII or not. As another example, it may be challenging for the two

[4] For evidence on incentives to report injuries to WC programs, see Biddle and Roberts (2003).
[5] These statistics are not directly comparable to the SOII estimates reported above for a variety of reasons. For example, an injury requiring one day away from work is in-scope for SOII but may not be included in the WC report, as it typically will not satisfy the waiting period requirement, see **www.dwd.state.wi.us/wc/WC_Basic_Facts.htm#WC_Claim_and_Indemnity_Information** (visited June 20, 2010).

sources to date cases in a comparable manner as the SOII intends to record an incident date associated with missed workdays relatively soon after the case occurs, whereas WC systems track other features of cases, including adjudication, medical costs incurred, and payment streams. The limited reporting time frame in the SOII is by design, so that estimates are produced in a timely fashion.

3. Methods

3.1 Linkages

Estimating injury and illness case counts requires a linkage between the two data sources, so that one can determine which cases are detected in both systems and which cases are detected in one system but not the other (cases unobserved in either system are imputed, as described below).

We use linked data for Wisconsin from Boden and Ozonoff (2008), which describes the dataset construction in detail. Their general strategy is to start with the SOII cases; determine via linking which SOII cases are also in the WC data; identify WC cases that are from employers in the SOII sample but not detected in SOII; and finally, restrict the collection of cases to reflect a common scope. This process effectively augments the cases observed in SOII with the additional cases from the same employer that appear in WC data. SOII sampling weights are used to calculate estimates of interest, such as case totals and the fraction of all cases detected by either of the two data sources.

Boden and Ozonoff apply a deterministic case record linkage first, with a minimum number of identical fields in both datasets implying a definite link. For remaining injuries, Boden and Ozonoff link cases using probabilistic record linkage (Fellegi and Sunter (1969); Belin and Rubin (1995)). Aside from linking cases, Boden and Ozonoff also link employers (and a linked case is presumed to imply an employer match as well). The employer linkage is necessary to drop WC unlinked cases for employers not in the SOII sample.

Note however that whereas the SOII data come from establishments chosen for the sample, the WC data tend to reflect reporting by firms. Consequently, the WC data are not detailed enough to allow one to consistently determine where within firms an injury occurs. This poses a methodological problem where employers have multiple establishments, only some of which are sampled by the SOII. In such situations we do not know whether an unlinked WC case occurred in a sampled portion of the employer. If occurring in a sampled establishment, it should be counted as a SOII missed case, otherwise it should be excluded from the analysis. The Boden and Ozonoff study recognizes this issue and makes a statistical adjustment to correct for it. Nevertheless, because the issue is an important one, we present our main results for single-establishment firms.

After linking the two data sets, Boden and Ozonoff impose several exclusion restrictions to limit the collection of cases to a common set of sectors and circumstances. This, generally speaking, restricts the data to include private sector activities where the SOII surveys establishments rather than relying on administrative censuses, and where the WC system requires coverage. In addition, BLS unlinked cases appearing not to satisfy the WC waiting period are excluded, as are WC unlinked cases not within the date subsampling range for the SOII establishments where case subsampling occurs.

Table 1 reports the resulting sample statistics for the 1998-2001 period, for single-establishment employers and for all employers, after the relevant exclusions are made. The single-establishment employer sample contains nearly 6000 establishments, with about 33,500 unweighted cases which represent over 130,000 cases using the SOII sampling weights. Of these, 52.6 percent are found in both data sources, 18.2 percent are found in SOII only, and 29.3 percent are found in WC only. The SOII capture rate is smaller in the whole sample than in the single-establishment subset.

Table 1: Sample Statistics

		Single-establishment employers	All employers
Case Captured by	SOII and WC	.526	.478
	SOII only	.182	.170
	WC only	.293	.353
Sample Size	Unweighted cases	33,541	73,614
	Weighted Cases	131,801	232,785
	Establishments	5956	9312

Notes. Data are for Wisconsin, 1998-2001. Sample means are weighted, using SOII sampling weights.

Although not reported in the table, there are other differences across the two samples, notably in the industry and employment size distributions. For example, the single-establishment employers tend to have smaller establishment employment and are more likely to be in the construction industry, compared with the whole sample.

3.2 Estimation

Our data include indicators for whether the case was detected in WC only, or in BLS only, or in both data sources (three possible outcomes). We estimate the probability of these outcomes as functions of explanatory variables with a multinomial logit (MNL) model,

$$(1)\quad p_{ij} = \Pr(Y_i = j) = \frac{\exp(X_i\beta_j)}{\sum_{k=1}^{3}\exp(X_i\beta_k)} \qquad j = 1,2,3$$

where i indexes the individual case, X_i refers to covariates, and j refers to the three possible outcomes. Estimation of equation (1) is on various samples of the linked data described above, using SOII sampling weights. Explanatory variables include characteristics of the establishment (industry, employment size class), the worker (age, gender) and the case itself (such as the number of days away from work and the date and the nature of injury or illness). The estimated probabilities may be combined to give the probability that one source detects the case; for example if j=1 refers to SOII unlinked cases and j=2 refers to linked cases, then (p_{i1} + p_{i2}) is the probability that SOII detects the case.

In our results below we show the marginal effect on the predicted probabilities of changing a covariate. We generate these by varying covariates X along one dimension, while holding X fixed at baseline values for all other components. For example, we compare predicted probabilities for different months of injury versus a baseline value (April), while holding fixed industry, year, and all other components of X at their

baseline values. The resulting changes in predicted probabilities estimate how capture rates vary over the course of the year, holding fixed other factors.

As in simple regression models, the real meaning of the marginal effects is often open to interpretation. Estimated effects may be true causal effects, or correlational effects only, they can be influenced by measurement error, and so forth. It is worth noting that estimated marginal effects can reflect non-sampling errors in the linkage process, or in scope exclusions, or in covariates' definitional incompatibility across sources. For example, if the WC system always miscodes a worker's age by adding 10 years, then the SOII would falsely appear more likely to miss older than younger workers. For a more realistic possibility, if short duration cases are mistakenly kept in the SOII data source despite being non-compensable in WC, then short duration can misleadingly indicate a failure of WC to capture the cases.

3.3 Imputing Cases Missed by Both Data Sources

The model in equation (1) is conditional on observing the case. One expects that some cases will be unobserved in both data sources. We account for this via capture-recapture analysis, under a baseline assumption that the WC and SOII data sources are independent. Capture-recapture analysis is a technique frequently used in epidemiologic and other settings where list comparisons are prominent; for background see Alho (1990), Alho et al (1993), Hook and Regal (1995), Tiling and Stearne (1999) and Chao et al (2001). To illustrate, without covariates (X_i is a constant in equation (1)) imagine a 2x2 table of probabilities as in Exhibit 1.

Exhibit 1: Source Capture Probabilities

Probability that a case is:	Not in WC	In WC
Not in SOII	p_0	p_3
In SOII	p_1	p_2

The top left cell gives the probability the case is entirely missed, p_0. The other three cells give probabilities that one or both sources capture the case. The MNL model in equation (1) identifies the relative sizes of the probabilities p_j (j=1,2,3), but not their sum. Assuming source independence, the probability that a case is missed by SOII is the same whether or not the case is observed in WC. Then the probability an injury occurs and is not observed in either surveillance system, p_0, can be determined by setting $p_0/ p_1 = p_3/p_2$. It is straightforward to then adjust the MNL predicted probabilities to return estimates for the p_j (j=1,2,3). When equation (1) is estimated with covariates the procedure is analogous, with observation-by-observation adjustments to the MNL predicted probabilities yielding estimates for the p_{ij}, where i indexes case observations. [6]

Positive source dependence would cause the capture-recapture methods to underestimate the amount of underreporting (Hook and Regal (1995), Brenner (1995)). Without additional sources of injury data or assumptions about the underlying distributions, we cannot determine the extent of such bias. A priori one might expect positive source dependence, so that cases more likely to be missed by one system are also more likely to

[6] The capture-recapture adjustments should be viewed as rough extrapolations and they are not without controversy in other settings (for example see Breiman (1994) and cites therein). We adopt the independence assumption because it is a common baseline in the literature on the undercount.

be missed by other systems. Although source dependence can affect the magnitude of estimates, it will generally not affect their direction.

3.4 SOII capture of WC cases

In an additional analysis, we restrict attention to cases in the WC data, and ask what case characteristics affect detection by SOII. This is a logistic analysis analogous to the model in equation (1), but conditional on WC reporting the case. We conduct this separate analysis because covariates available only in the WC data may improve our understanding about why cases reported to WC are not reported to SOII.

4. Results

4.1 MNL Estimation of Capture Propensities

Table 2 summarizes the main results from the multinomial logit model estimation of equation (1). The estimation sample is restricted to employers with only one establishment. The first two rows give overall probabilities, while the remaining rows show how predicted probabilities from the model change with changing covariates. Standard errors are generated using bootstrap techniques.[7] SOII capture is the proportion captured by both systems, plus the proportion captured only by SOII ($p_1 + p_2$ in the 2x2 table above). WC capture is the proportion captured by both systems, plus the proportion captured only by WC ($p_2 + p_3$).

The first row gives raw probabilities, derived by estimating equation (1) without covariates and then deriving the proportion reported to neither system using the assumption of source independence. The unadjusted analysis suggests that the SOII captures about 64 percent of all days away from work cases for this population. WC detects a greater fraction of cases, but still appears to miss a substantial number. These overall probabilities are roughly similar to those reported in Boden and Ozonoff (2008). The second row of table 2 reports predicted probabilities at baseline covariates. Although not reported separately, the sample averages for the model-predicted probabilities are virtually the same as the raw probabilities reported in the first row.

The remaining rows of table 2 give covariate means (the last column), and covariate marginal effects on the probabilities that the respective data sources capture the case. For example, the number -.075 for age group 16-19 means that the model predicts the SOII capture rate is 7.5 percentage points lower for a case involving a 16-19 year old than a case involving a 35-44 year old (the baseline age group), holding other covariates fixed at baseline values. The -.135 figure in the third column indicates that the analogous WC capture probability differential is 13.5 percentage points. This implies less data source overlap for this age group than for the others; these cases are less likely to be linked cases. The final column reports that in 3.5 percent of cases the worker is 16-19 years old.

[7] Replicate samples of establishments are drawn within strata defined by industry and employment size class (this is roughly consistent with the strata definitions of the SOII). Certain parts of the data construction, notably the weights adjustments for multi-establishment employers, are not repeated for each replicate.

Table 2: Factors Associated with Case Detection in the SOII and in WC, Single-Establishment Firms

		SOII capture		WC capture		
		Prob.	Std. error	Prob.	Std. error	
Raw probability		.643	.010	.743	.007	
Predicted at baseline		.686	.036	.590	.043	
Covariate		Marginal effect	Standard error	Marginal effect	Standard error	Sample means
Age group	16-19	-.075	.045	-.135	.048	.035
	20-24	.015	.025	-.013	.025	.113
	24-34	.009	.017	-.015	.024	.239
	35-44	0	-	0	-	.306
	45-54	-.023	.020	.032	.025	.204
	55-64	-.056	.022	.004	.037	.093
	65 +	-.058	.046	.015	.060	.011
Female		-.051	.022	.010	.020	.269
Part of Body	head, neck	-.057	.034	-.070	.066	.052
	up. extremities	-.045	.019	.147	.026	.267
	trunk	0	-	0	-	.393
	lo. extremities	-.002	.019	.072	.025	.195
	internal organs	-.100	.087	-.546	.039	.008
	other	-.008	.026	.013	.038	.085
Nature	sprains, strains	0	-	0	-	.508
of Injury	other traumatic	.059	.016	-.201	.022	.279
	CTS, tendonitis	-.165	.034	-.291	.035	.048
	other disorders	-.026	.018	.334	.033	.140
	miscellaneous	.071	.034	-.077	.067	.025
Month	Jan	-.091	.038	-.068	.035	.094
of Injury	Feb	-.033	.038	-.060	.036	.084
	March	-.026	.038	-.061	.042	.088
	April	0	-	0	-	.078
	May	-.016	.036	-.043	.045	.085
	June	-.058	.035	-.076	.041	.091
	July	-.023	.036	-.021	.035	.090
	Aug	-.004	.033	.020	.031	.088
	Sept	-.025	.035	-.012	.045	.075
	Oct	-.018	.034	-.010	.037	.085
	Nov	-.066	.039	-.025	.040	.071
	Dec	-.136	.034	-.018	.042	.071
Duration	1 (0-6 days)	0	-	0	-	.392
(weeks to	2	.135	.018	.275	.027	.178
return	3	.120	.021	.282	.026	.091
to work)	4	.124	.026	.300	.032	.055
	5	.131	.023	.307	.037	.045
	6	.158	.024	.314	.039	.032
	7	.114	.036	.323	.038	.028
	8 +	.139	.018	.318	.033	.179

Notes. The sample includes injury and illness cases from single-establishment employers in Wisconsin, 1998-2001 (N=33,541). Estimation uses SOII sampling weights. The model pseudo-R^2 is 0.103. Raw probabilities shown in the first row of numbers have no covariate controls; other

results are derived from a multinomial logit model. The "predicted at baseline" row gives predicted probabilities at covariate values corresponding to omitted groups. The "marginal effect" columns report changes in the relevant predicted probability varying values for one covariate, holding other covariates fixed at baseline values. The final column gives the sample distribution across group values. Year effects are included as controls.

Age and gender are generally not important predictors of capture rates. However, there is some evidence that women have a slightly lower probability than men of having cases reported to SOII. The SOII appears to do a slightly worse job of capturing workers at either end of the age distribution (as compared to workers aged 20-54), and as mentioned above there is more of a tendency for either system to miss cases of very young workers.

Case characteristics such as nature of injury, part of body, and month of injury appear to be somewhat more important predictors. For example, both the SOII and WC are less likely to detect carpal tunnel and tendonitis cases than the typical case involving strains and sprains.[8] Unfortunately, the concordance between the coding schemes used in the WC and SOII are too coarse for further analysis here. We will present results for a more detailed set of nature of injury codes in our analysis of WC cases below. The SOII is more likely to miss a case if it occurs in December. This likely reflects the fact that the SOII data collection begins just after the close of the year, so end-of-year cases that are not immediately apparent may be missed by the SOII but not necessarily by WC. There appear to be other month effects, notably a January effect in the SOII.

Another interesting result involves shorter duration cases. Duration is measured in terms of the week of return to work, so duration=1 in these tables means the case has 6 or fewer days away from work. Cases where the worker returns to work within one week have much lower probabilities of capture by either data source. The lower capture rates for the SOII follow from a group of cases that are coded as having zero days of temporary disability in the WC system. These cases may be adjudicated cases, which may have initially been considered as not work-related and therefore not entered in the SOII data. At a later point, they may have been settled with a lump-sum payment with no indication of the number of days lost. Alternatively, they may not have been eligible for the SOII because the injured worker did not lose a full day of work after the date of injury. The lower capture rates for the WC data – a difference of roughly 30 percentage points - may reflect difficulty in determining whether a case meets the WC waiting period requirement.[9] Some cases appearing in SOII with short durations perhaps were mistakenly not excluded from the analysis as they may not have in actuality satisfied the WC waiting requirement. Or, workers may not recognize they are eligible for benefits or may not think it worth filing a claim for relatively small benefits amounts and therefore fail to initiate a WC claim (Azaroff et al. (2002)). Shorter duration cases account for almost 40 percent of the days away from work cases in this sample, and this is an important group of cases for future analysis.

The remaining covariates, not listed in the table, are characteristics of the SOII establishment. As a general rule their effects tend to be small and imprecisely estimated.

[8] Please note that the nature and part of body codes in the two sources are derived under different coding schemes. When cases are linked and both sources have valid values for a case characteristic, we default to using the WC measure, so that measures in the logit analysis below are derived from the same data source.

[9] The large effects here are the primary reason why the predicted baseline WC capture probability in the second row of table 2 is smaller than the WC raw probability shown in the first row.

In terms of industry effects, there is some evidence that the SOII is relatively less likely to capture a case if it is in Transportation, Communications and Utilities, or in Wholesale or Retail Trade than if it is in Agriculture, Construction or Nondurables Manufacturing. For WC cases, cases in the Construction industry have a relatively high predicted probability of detection. In terms of establishment size, cases in the largest establishments (by employment) appear to be more likely to be detected in WC and less likely in SOII, as compared to medium size establishments. There are no statistically significant differences depending on whether the establishment has cases subsampled, or whether the establishment is pre-notified (meaning, the establishment was exempt from OSHA record-keeping and therefore was provided OSHA log forms prior to the survey year)[10]. These are interesting negative results in that establishments with case subsampling might be expected to understand both reporting systems well. The results for case subsampling establishments do not immediately suggest measurement problems associated with excluding cases on the basis of injury date.

The above estimation is repeated for a broader set of SOII respondents; the sample now includes establishments that are part of multi--establishment employers. The raw SOII capture rates are smaller for this set, suggesting that the SOII more accurately captures cases in single-establishment firms. The covariate set is the same as that in above, except that now there are additional indicators for whether the establishment is a single-establishment firm or not.[11] Although not reported in the table, the estimated marginal effects are largely similar in the two samples. Table 3 shows that the SOII is much more likely to identify otherwise similar cases in single-establishment than multi-establishment firms or firms of unknown type. Whether this reflects real differences or instead reflects non-sampling errors in failing to delete some unlinked WC cases in multi-establishment firms is not known at this point. But whatever the source, it appears to operate for SOII capture rates only, as the WC capture rate does not vary with employer type.

Table 3. Case Detection by Employer Status, All Employers

		SOII capture		WC capture		
		Prob.	Std. error	Prob.	Std. error	
Raw probability		.575	.015	.738	.005	
Predicted at baseline		.592	.031	.553	.031	
		Marginal effect	Standard error	Marginal effect	Standard error	Sample means
Employer	Single-Estab	0	-	0	-	.566
	Multi-Estab	-.109	.021	.002	.016	.352
	Unknown	-.226	.045	-.026	.032	.082

Notes. The sample includes injury and illness cases from all employers in Wisconsin, 1998-2001 (N=73,614). Estimation uses SOII sampling weights. The model pseudo-R^2 is 0.106. Raw probabilities shown in the first row of numbers have no covariate controls; other results are derived from a multinomial logit model. The "predicted at baseline" row gives predicted probabilities at covariate values corresponding to omitted groups. The "marginal effect" columns report changes in the predicted probability varying employer status, holding other covariates fixed at baseline values. The final column gives the sample distribution across group values. The full set of controls include the characteristics in Table 2 and year effects.

[10] The SOII now pre-notifies all sampled establishments.

[11] We do not always know whether the establishment is a single-establishment employer as of the survey year. This is identified using the universe as of the survey year, and while all SOII establishments are of course in the universe at and just prior to sampling, they are not all in the universe as of the survey year.

4.2 Logistic Estimation of SOII Capture Probabilities

As noted above, we also take a further look at WC cases, and ask what type of WC cases the SOII is more or less likely to capture. Our main purpose here to look at the marginal effects of certain variables unique to the WC data. The WC data have more detailed nature of injury codes than were reported in tables 2 and 3 (as do the SOII data), and it is interesting to disaggregate these where feasible. One can also infer from the WC data whether the case entered the WC data in the year of injury, or later. This is of some interest given the suggestion above that late-year cases may come too late for SOII. Finally, there are some indicators in the WC data for what appear to be unusual WC cases, based on status codes maintained by the WC system, and whether the case has zero days of temporary disability payments (lost workday payments) associated with it.

We estimate this by restricting the sample to WC cases and estimating a logit analog to equation (1). The outcome of interest is whether the case is detected by the SOII. As before, we take as our main sample single-establishment firms and also show estimates for an alternative sample including multi-establishment firms (and firms of unknown type). The controls are similar to those in tables 2 and 3, and because the same patterns prevail we report only an abbreviated set of effects here in table 4.

Table 4 shows that the SOII tends to do a better job of capturing WC cases that are discrete identifiable events as opposed to latent or chronic injuries or illnesses. For example, amputations, fractures, injuries caused by contact with a foreign body, and cases with multiple physical injuries tend to be captured with higher probability than other cases in the SOII. On the other hand, the SOII predicted probability of capture is lower for cumulative injuries and certain occupational diseases.

In terms of the timing of cases, nearly 15 percent of WC cases come into the system after the year of injury. Many of these will not be reported in SOII, possibly because they were unknown or disputed at the time the SOII surveyed the establishment. Such cases are perhaps 20-30 percent less likely to be in SOII, even controlling for other characteristics such as nature and duration of injury. Furthermore, the November and December end-of-year effects disappear or even perhaps reverse in these models. That is, the SOII is at least as likely to capture December as April cases, but only provided they appear on the WC rolls before the year's end.

The other reported effects in table 4 point to some possible unusual situations or cases. The WC system reports a case status code, and about 97 percent of cases have a status code of "electronic" or "final." Cases marked as "final" have WC payment information included in the initial supplementary filing. Presumably, most of these cases have been provisionally recognized by the employer. Cases marked as "electronic" are those filed electronically; unfortunately, there is little else that this status flag reveals about cases. Of the remaining 3 percent of WC cases, SOII capture rates are noticeably smaller. The majority of these cases have the "award" status, indicating that a formal order has been written providing compensation for the claim. Cases with award status are typically disputed cases adjudicated in the claimant's favor or settled by the claimant and the employer's insurer. When a case is disputed, the final determination of whether the injury

or illness is work related can occur long after the year of injury and can result in a lump-sum payment without distinguishing the number of days away from work. [12]

Table 4. Factors Associated with Detection in SOII, for Cases Reported to WC

	Single-establishment employers		All Employers		
Covariate	Marginal effect	Standard error	Marginal effect	Standard error	Sample means
Nature of Injury					
<Reclassified codes>	-.004	.048	-.005	.045	.003
Amputation, severance	.146	.032	.132	.028	.010
Burn	.046	.043	.059	.030	.014
Concussion	.054	.074	.059	.051	.001
Contusion	.012	.019	.026	.016	.059
Crushing	.053	.036	.061	.027	.013
Dislocation	-.144	.126	-.073	.098	.007
Electric shock	.117	.123	.080	.105	.001
Foreign body	.144	.040	.130	.037	.004
Fracture	.066	.025	.072	.019	.072
Hernia, rupture	.031	.025	.008	.021	.030
Infection	.081	.053	.119	.040	.001
Inflammation	-.066	.036	-.043	.025	.012
Laceration	.042	.019	.057	.017	.053
Puncture	-.036	.047	.009	.039	.008
Sprain	.041	.018	.041	.020	.049
Strain	0	-	0	-	.465
Other specific injuries	-.008	.017	-.008	.013	.113
Respiratory disorders	.075	.078	.038	.055	.001
Dermatitis	.157	.064	.192	.053	.003
Other occup. disease	-.397	.183	-.422	.106	.001
Hearing loss	-.439	.174	-.565	.117	.008
Mental stress	-.097	.161	-.022	.127	.001
Carpal Tunnel Syn.	-.061	.037	-.033	.027	.029
Other cumulative inj.	-.107	.040	-.108	.030	.019
Mult. injuries, physical	.092	.024	.095	.024	.022
Mult. injuries, physical and psychological	-.164	.155	-.117	.141	.001
Report to WC after injury year	-.292	.030	-.221	.022	.147
Month of Injury					
Jan	-.074	.032	-.055	.022	.094
Feb	-.028	.028	-.025	.022	.083
March	-.025	.033	-.012	.021	.085
April	0	-	0	-	.079
May	-.012	.027	-.005	.021	.084

[12] Over 90 percent of award-status cases have zero days of temporary total disability (TTD) recorded. A TTD day is roughly comparable to a lost workday in the SOII (however, cases with 1-3 lost workdays can be coded as having zero TTD days due to the WC waiting period and settled cases may not report lost workdays as part of the settlement agreement). The status code "no lost time" indicates the case was coded as having no lost workdays in an initial supplementary report; the majority of these cases subsequently involve lost workdays.

June	-.033	.029	-.035	.021	.088
July	-.005	.025	.010	.020	.091
Aug	.011	.025	-.005	.020	.090
Sept	.004	.025	.003	.021	.075
Oct	.019	.024	-.010	.020	.085
Nov	.025	.026	-.004	.021	.072
Dec	.055	.024	.027	.020	.075
TTD days = 0 indicator	-.299	.033	-.285	.023	.121
Duration (week return to work)					
1 (0-6 days)	0	-	0	-	.330
2	.019	.016	.017	.012	.191
3	.007	.019	.008	.013	.100
4	.004	.024	.018	.020	.061
5	.014	.021	.011	.017	.050
6	.044	.021	.016	.017	.035
7	.014	.031	.011	.022	.032
8 +	.034	.015	.018	.012	.201
WC status code					
Award	-.199	.058	-.155	.042	.024
Electronic	-.054	.022	-.002	.018	.164
Final	0	-	0	-	.801
No Lost Time	-.202	.068	-.118	.047	.009
<Reclassified Other>	-.268	.178	-.135	.118	.002
N	27,476		62,941		
Pseudo-R^2	0.126		0.115		

Notes: The first set of estimates is for single-establishment employers. The second set of estimates includes single- and multi-establishment employers. The final column gives the sample distribution across group values, for the single-establishment sample. Year, gender, age category, industry, establishment employment, part of body, case subsampling and prenotification effects are included as controls but not reported. The model for all employers additionally includes but does not report controls for whether the employer has one or more than one establishment.

More strikingly, cases with zero days of temporary total disability (TTD days) are nearly 30 percent less likely to be captured in SOII than are otherwise comparable cases. Such cases, which make up about 12 percent of the sample, may reflect settlements or adjudicated cases. (Because workers' compensation does not pay temporary disability benefits during the first 3 days lost from work, they could also be cases with 1-3 lost workdays). Taken together, these results suggest that subjective opinions about whether an injury or illness is really a bona fide work-related case can affect undercount estimates.

5. Interpretations and Conclusions

This study documents characteristics of work-related injuries and illnesses which make them more or less likely to be captured by the SOII or by the WC system. The results in the previous section largely fall into one of three categories: reporting issues, methodological issues, and timing issues.

Reporting issues. Employees might be less likely to make a WC claim for a work-related injury if the burden of reporting is greater than the financial incentives for doing so. Workers in these categories appear to include workers under age 20, who may be less

knowledgeable or assertive about reporting injuries or who may have lower wages and therefore less potential for income replacement. Similarly, workers are only compensated for wages for the fourth and greater days of lost work if their injury or illness keeps them out of work for less than a week. The results confirm that the capture rate for both the SOII and WC is lowest for cases with disability lasting less than a week. In some adjudicated cases, the employer may doubt work-relatedness and therefore is unlikely to have recorded the case in the OSHA log in a timely manner.

A related set of cases are those which are systematically unreported to both systems. While it is not possible to characterize these cases in our regression analysis, there are a number of worker and employer incentives which might cause a case to go unreported. Workers may avoid reporting injuries because they may believe that reporting will lead to retaliation by their employers. Alternatively, some employers have group safety incentives in place, leading to pressure by fellow workers to avoid reporting. Others may simply not know that they should report or may not want to bother reporting. Employers may not want to acknowledge injuries because of concerns about increased workers' compensation costs or because injury rates are a factor in determining who receives contracts. Finally, workers' compensation systems may erect barriers to receipt of benefits for some classes of injuries, which employers may then not report to the SOII. (See Azaroff et al. (2002), Boden and Ruser (2003).) Given the many potential reporting issues, it is difficult to imagine a simple fix to capture these types of cases. However, they explain only a portion of the undercount.

Methodological issues. A second potential explanation for the estimated undercount are a variety of methodological and data quality issues. The strongest example for the Wisconsin data is the difference in results for single versus multi-unit establishments. While it is certainly possible that multi-unit establishments have different reporting rates than single unit establishments, the size of the difference is large and only affects the SOII capture rate and not the WC capture rate. This suggests that our attempts to take into account the differences in reporting unit between workers' compensation and the SOII may have been inadequate. Progress can be made to determine the potential size of this issue if future work entails WC data with sufficiently detailed information on the establishment in which the injury occurred. Another potential source of methodological issues involves restricting the scope of the SOII and WC data to a common universe. For example, the SOII needs to be restricted to cases that meet the WI WC waiting period, and similarly WC needs to be restricted to cases with at least one day away from work.[13] While some of these issues can be remedied by a careful reading of the relevant WC and OSHA regulations, some ambiguities will likely remain. One can estimate the bounds on the potential impact of these issues with a sensitivity analysis. We leave this for future research.

Timing issues. Also, many discrepancies in cases reported to either WC or the SOII and not the other source appear to stem from timing issues. The SOII collects data in the first six months of the year following the date of injury, while the WC system is continually updated with new information. BLS could potentially improve the SOII capture rate by collecting information for a longer period of time but doing so would likely affect the timeliness and accuracy of the published statistics. The SOII also appears to do a poor job in capturing cases in which the day of injury might differ from the first day of lost work, with, for example, carpal tunnel syndrome cases. For these cases, further information

[13] See Oleinick and Zaidman (2010) for more complete information.

from additional sources such as medical records might help to predict cases that are difficult to capture. However, this type of case by case research would be difficult to replicate at a national level as is necessary for the SOII. As is, the BLS data should be interpreted as recording injuries known early in the year following injury.

Acknowledgements

This work is supported in part by NIOSH grant 5 R01 OH 007596 (Boden). Opinions presented in this paper are those of the authors and do not represent the opinions or policies of the Bureau of Labor Statistics or any other agency of the U.S. Department of Labor.

References

Alho, JM (1990). Logistic regression in capture-recapture models. Biometrics 46(3):623-635.

Alho JM, Mulry MH, Wurdeman K, Kim J (1993). Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. Journal of the American Statistical Association 88(423):1130-1136.

Azaroff LS, Levenstein C, Wegman DH. (2002). Occupational injury and illness surveillance: Conceptual filters explain underreporting. American Journal of Public Health 92(9):1421-1429.

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.

Biddle J and Roberts K (2003). Claiming Behaviour in Workers' Compensation. Journal of Risk and Insurance, 70(4):759–780.

Boden LI, Ozonoff A (2008). Capture-Recapture Estimates of Nonfatal Workplace Injuries and Illnesses. Annals of Epidemiology 18(6): 500-506.

Boden LI, Ruser J. (2003). Choice of Medical Care Provider, Workers' Compensation "Reforms," and Workplace Injuries. Review of Economics and Statistics. 85(4):923-929.

Breiman, Leo (1994). The 1991 Census Adjustment: Undercount or Bad Data? Statistical Science 9(4): 458-475.

Brenner, Herman (1995). Use and Limitations of the Capture-Recapture Method in Disease Monitoring with Two Dependent Sources. Epidemiology 6(1):42-48.

Chao A, Tsay PK, Lin S-H, Shau W-Y, Chao D-Y (2001). The Applications of Capture-Recapture Models to Epidemiological Data. Statistics in Medicine. 20(20):3123-3157.

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

Hook EB, Regal RR (1995). Capture-recapture methods in epidemiology: methods and limitations. Epidemiologic Reviews 17(2):243-64.

Oleinick A , Zaidman B (2010). The need for data harmonization - a response to Boden and Ozonoff. American Journal of Industrial Medicine 53(8): 854-855.

Rosenman KD, A Kalush, MJ Reilly, JC Gardiner, M Reeves, Z Luo (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, John W. (2008). Examining evidence on whether BLS undercounts workplace injuries and illnesses. Monthly Labor Review August, 20-32.

Tilling K, Sterne JAC (1999). Capture-recapture models including covariate effects. American Journal of Epidemiology. 149(4):392-400.