Capture-Recapture Estimates of Workplace Injury Rates:
Final Progress Report

NIOSH Grant R01 OH007596
Start date: 9/1/2002
End date: 8/31/2006

Principal Investigator: Leslie I. Boden, Ph.D.
Department of Environmental Health
Boston University School of Public Health
715 Albany Street
Boston, MA 02118
Email: lboden@bu.edu

Co-Investigator: Alexander Ozonoff, Ph.D.
Department of Biostatistics
Boston University School of Public Health
715 Albany Street
Boston, MA 02118

TABLE OF CONTENTS

List of Figures 2
List of Tables 2
List of Abbreviations 2
ABSTRACT (Including Significant Findings) 3
SIGNIFICANT FINDINGS 3
TRANSLATION OF FINDINGS 4
OUTCOMES/RELEVANCE/IMPACT 5
SCIENTIFIC REPORT 5
BACKGROUND 5
DATA AND METHODS 8
RESULTS 14
SUMMARY OF SPECIFIC AIMS OF THE GRANT AND SUCCESS IN MEETING THEM 17
PUBLICATIONS 18
Inclusion of Gender and Minority Subjects 18
Inclusion of Children 18
Materials Available for other Investigators 18

List of Figures

No figures are used in this report.

List of Tables

Table A. Estimates of Reporting Completeness for 6 States: Lost Work Longer than Workers' Compensation Waiting Period

Table B. Estimates of Reporting Completeness for 6 States: Lost Work Longer than Workers' Compensation Waiting Period. Odds Ratio = 5

Table 1. Information about State Workers' Compensation Data

Table 2. Two-source Capture-Recapture Model

Table 3. Estimates of Reporting Completeness for 6 States: Lost Work Longer than Workers' Compensation Waiting Period

Table 4. Estimates of Reporting Completeness for 6 States: Impact of Assumptions about Source Dependence

List of Abbreviations

BLS: Bureau of Labor Statistics, U.S. Department of Labor
OSHA: U.S. Occupational Safety and Health Administration
PPD: permanent partial disability
SOII: BLS Annual Survey of Injuries and Illnesses
TTD: temporary total disability

ABSTRACT (Including Significant Findings)

This study has used the two most widespread and comprehensive sources of nonfatal occupational injury and illness reporting to evaluate completeness of reporting and factors that affect differential reporting. These sources are state workers' compensation data and the Survey of Occupational Injuries and Illnesses (SOII) conducted annually by the Bureau of Labor Statistics (BLS).

For 1998-2001, the study compares reported injuries from these two sources by matching individual lost-time injuries and illnesses in 6 states: Minnesota, New Mexico, Oregon, Washington, Wisconsin, and West Virginia. It uses probabilistic matching methods to link data from the two sources. It then uses capture-recapture analysis for each of these states to develop improved measures of nonfatal injury incidence. The study also uses logistic regression to account for differential capture by employer, injury, and worker characteristics. This helps to identify factors associated with underreporting. Then, we also compare the degree to which the two data systems capture occupational injuries and illnesses in the states studied.

Using an assumption that biases estimates of underreporting toward zero (that reporting from each source is independent of the other), we find substantial underreporting in both workers' compensation and the BLS survey, with the exception of reporting of lost-time cases in Washington State and West Virginia. The state median of estimated underreporting in workers' compensation is 25%, and the median in the BLS survey is about 40%. The range in workers' compensation is from 5% to 37%, and the range in the BLS survey is from 23% to 47%.

We tested the sensitivity of these estimates to the assumption of source independence. If we assume that the odds of a case's being reported to a workers' compensation system are 5 times as high if it is reported to the BLS than if it is not, then the state median of estimated underreporting in workers' compensation is 45%, and the median in the BLS survey is about 50%.

We also found that simply counting the cases that appear in either system substantially improves the overall completeness of reporting in most states. On the assumption of source independence, the minimum ascertainment in any state is 86%.

An important limitation of this study is that injuries or illnesses that employers are unaware of will be absent from both the workers' compensation and BLS survey data. This may happen because workers are concerned about the stigma of claiming an injury because of concern about employer retaliation, or because the worker and employer are unaware of the work-relatedness of the condition. This makes capture-recapture analysis particularly inappropriate for determining underreporting of chronic occupational diseases.

SIGNIFICANT FINDINGS

Reporting rates. Using an assumption that biases estimates of underreporting toward zero (that reporting from each source is independent of the other), we find substantial underreporting in both workers' compensation and the BLS survey, with the exception of reporting of lost-time cases in Washington State and West Virginia. The state median of estimated underreporting in workers' compensation is 25%, and the median in the BLS survey is about 40%. The range in workers' compensation is from 5% to 37%, and the range in the BLS survey is from 23% to 47% (Table A).

Table A. Estimates of Reporting Completeness for 6 States: Lost Work Longer than Workers' Compensation Waiting Period

Percent reported to:	WA	WV	OR	WI	NM*	MN
Workers' Compensation	94%	91%	78%	75%	67%	64%
Bureau of Labor Statistics	51%	75%	55%	65%	50%	71%

Either	96%	97%	88%	92%	84%	88%

Note: Calculations assume source independence.
*New Mexico has a 7-day waiting period. The other states have 3-day waiting periods.

We tested the sensitivity of these estimates to the assumption of source independence. If we assume that the odds of a case's being reported to a workers' compensation system are 5 times as high if it is reported to the BLS than if it is not, then the state median of estimated underreporting in workers' compensation is 45%, and the median in the BLS survey is about 50% (Table B).

Table B. Estimates of Reporting Completeness for 6 States:
Lost Work Longer than Workers' Compensation Waiting Period
Odds Ratio = 5

Percent reported to:	WA	WV	OR	WI	NM*	MN
Workers' Compensation	83%	85%	61%	61%	45%	52%
Bureau of Labor Statistics	49%	71%	44%	53%	37%	55%
Either	86%	91%	68%	74%	57%	68%

*New Mexico has a 7-day waiting period. The other states have 3-day waiting periods.

We also found that simply counting the cases that appear in either system substantially improves the overall completeness of reporting in most states. On the assumption of source independence, the minimum ascertainment in any state is 84% (Table 1).

Variation in Reporting Rates by Covariates. The coefficients of the multinomial logistic regression that estimates the probability of an injury's being in both sources, just the SOII, or just the state workers' compensation data are not directly interpretable, so we take the approach of comparing the predicted values of reporting completeness for alternate employer, injury, and worker characteristics for each state. We begin with a base case: a back strain that occurred in a manufacturing establishment with 100-250 workers to a male worker aged 35-44. We then predict the effect of changing covariates one at a time.

There is little state-to-state consistency in the impact on ascertainment of many of the factors we examined. However, some alterations from the base case produce changes of more than 5 percent in the probability of ascertainment for more than half the states.

Differential ascertainment for specific characteristics, occurred mostly in the BLS survey. In all 6 states, the smallest firms, those with 11-49 employees reported at least 5 percent more than did the mid-sized firms in the base case. Wholesale and retail firms had lower ascertainment than manufacturing firms in 5 of 6 states. In addition, upper extremity cumulative trauma disorders had lower rates of ascertainment in 5 of 6 states. In 3 of the states, ascertainment was much lower (38% to 70% lower). Chronic occupational disease was reported so infrequently that ascertainment could not be estimated in 3 of 6 states. Even where estimates were made, we do not think they are reliable.

The only factor affecting workers' compensation injury ascertainment in more than half the states is age. In the age 55+ group, injury ascertainment is at least 5% higher in 4 of the 6 states.

TRANSLATION OF FINDINGS

This research raises important questions about the accuracy of both workers' compensation and BLS data on workplace injuries and illnesses. It also raises concerns about the adequacy of workers'

compensation benefits, over and above those raised by previous studies. Earlier studies of the adequacy of workers' compensation benefits showed low replacement of lost earnings, but these studies did not account for the fact that many injured workers seem to get no workers' compensation benefits at all.

These two systems are the only national U.S. injury and illness surveillance systems. Data from these systems can be used to establish priorities for investing in occupational safety and health and to establish prevention priorities among injury categories and among industries. To the extent that the data are inaccurate, their value in establishing priorities is reduced.

The implications of this study include:

(1) The level of ascertainment of the BLS SOII is poor, and steps should be taken to identify the reasons for this and then to improve the accuracy of the BLS survey.
(2) Many workplace injuries and illnesses go uncompensated in some state workers' compensation systems. Other systems seem to compensate a relatively high proportion of injuries and illnesses. Steps should be taken to determine the reasons for these differences (in addition to different waiting periods) and to provide this information to the states.

OUTCOMES/RELEVANCE/IMPACT

Two outcomes of this project are of note. First, the BLS has decided to work with the PI of this grant to try to understand the reasons for the low ascertainment of its SOII. Second, the State of California has asked the PI to work with it to determine the completeness of reporting to the California workers' compensation system. If we find substantial underreporting, the state plans to take steps to understand the reasons for underreporting and to attempt to achieve more complete reporting.

SCIENTIFIC REPORT

BACKGROUND

Injury and illness reporting systems provide the basis for determining the level and allocation of resources to the prevention of occupational injuries and illnesses. Reported injury rates also are used to measure the success or failure of prevention efforts. Yet, researchers and policymakers sometimes treat *reported* injury rates as if they were *actual* injury rates, and they often do not consider possible disparities between the number of injuries and the number of reported injuries.

The most commonly used sources of state and national estimates of injury frequency are state workers' compensation data and the Survey of Occupational Injuries and Illnesses (SOII) conducted annually by the Bureau of Labor Statistics, U.S. Department of Labor. Most state workers' compensation agencies collect first reports of injury (and many collect subsequent information). Frequently, the states themselves or organizations involved in the process of setting workers' compensation rates publish information on the number and type of injuries incurred. The BLS collects information from a statistical sample of establishments on injuries that the Occupational Safety and Health Administration (OSHA) requires employers to record.

There is widespread evidence of substantial disparities between the number of injuries that are reported and the actual number that occur. This has several implications. First, funding for research and prevention may be less than it should be because policymakers think that the problem is smaller than it actually is. Second, reporting may be particularly incomplete for specific conditions, subpopulations, and employer types. As a consequence, injuries and illnesses for which we have particularly poor measures of incidence may receive inadequate public health attention.

A series of studies beginning in the 1980s demonstrated that both nonfatal and fatal injuries are underreported. Hanrahan (1987) reports an overall injury rate in Wisconsin that is 31 percent higher using BLS data than using workers' compensation data. Some or all of this disparity might be caused by the fact that Wisconsin has a 3-day waiting period for income benefits (a threshold for entry into the workers'

compensation database) while the BLS SOII does not. Hanrahan could not access BLS data at the injury level, so he could not account for this. Still, 8 percent of workplaces reported more injuries involving lost workdays to workers' compensation than they reported to the BLS.

A study of OSHA injury reporting (Seligman, Sieber, Pederson, Sundin, and Frazier 1988) showed that only 75 percent of establishments required to keep OSHA 200 injury and illness logs actually did so. 60 percent of companies employing between 11 and 99 workers kept these records, while 95 percent of companies with 500 or more employees kept them. This shows underreporting on the forms used in the BLS SOII, and it demonstrates underreporting is greater among smaller firms.

Another study of OSHA injury reporting attempted to determine not only the proportion of employers that actually kept logs but the accuracy of the logs that were kept (Eisenberg and McDonald 1988). Of the 192 establishments visited in this study, almost 90 percent kept the OSHA 200 log, but the accuracy of these logs was called into question when the study attempted to reconstruct the log from employee medical records, clinic logs, company and insurer accident reports, accident and health benefit reports, and absentee records. Using this reconstruction as a measure of "correct" reporting (although some injuries were probably missed in the reconstruction), the study estimated that the OSHA logs under-recorded lost workday cases by about 25 percent and over-recorded cases not involving lost workdays by about 10 percent. Part of the under-recording of lost workday cases involved classifying these cases as not involving lost workdays.

Fine, Silverstein, Armstrong, Anderson, and Sugano (1986) conducted a study of surveillance sources for upper extremity cumulative trauma disorders at three large automobile manufacturing plants, two of which had recently designed improved cumulative-trauma surveillance systems within their occupational medicine departments. They compared four data sources: the OSHA injury logs used in the BLS SOII, workers' compensation cases, records of medical absences longer than three days, and plant medical records. For acute traumatic cases, injury logs and medical absence records generated similar incidence rates, workers' compensation rates were substantially lower, and rates derived from plant medical records were four times as high as from injury logs. For cumulative trauma disorders, medical absence and plant medical records produced incidence rates an order of magnitude higher than workers' compensation and two orders of magnitude higher than injury logs. The increasing recognition of cumulative trauma disorders in the past two decades has undoubtedly improved reporting.[1] However the size of the disparity, even given this possibility, is disquieting.

A more recent study of the automobile manufacturing industry in 1986-1989 compared incidence rates for musculoskeletal disorders using the OSHA 200 log (the source of the BLS SOII data), workers' compensation, sickness and accident insurance, and self-reported symptoms, medical treatment, and lost time (Silverstein, Stetson, Keyserling, and Fine 1997). Of particular relevance to the proposed study is variation from plant to plant and by calendar year in the relationship between incidence rates calculated from the OSHA 200 logs and from workers' compensation data. Indeed, in some cases the OSHA incidence rate was much lower than the workers' compensation rate, and in other cases it was considerably higher. This finding suggests that the two sources provide different kinds of information about injuries.

A study in the semiconductor industry examined whether injury cases classified by employer medical clinics to be occupational in origin were recorded on the injury log (McCurdy et al. 1991). It found that the logs contained only 60 percent of cases that met the BLS reporting criteria.

Several more recent studies have focused on the filing of workers' compensation claims. Biddle, Roberts, Rosenman, and Welch (1998) compare legally required physician reports of occupational diseases with workers' compensation claims in Michigan. The Michigan definition of occupational disease (as distinguished from occupational injury) includes many musculoskeletal disorders and, indeed, over 36

[1] Between 1982 and 1992, the BLS reported an increase in reported cumulative trauma disorders from 22,600 to 332,100 cases (Panel on Musculoskeletal Disorders and the Workplace 2001, p. 37). Much of this increase may have been a consequence of more complete reporting.

percent of reported diseases were for sprains, strains, or carpal tunnel syndrome. Biddle et al. found that less than half (and perhaps as little as 10 percent) of workers with known or suspected occupational diseases had filed workers' compensation claims. In a recent follow-up study focusing on musculoskeletal disorders, Biddle and Roberts (2001) concluded that, among those who missed more than 7 days from work (making them eligible for workers' compensation income benefits), less than 60% filed for these benefits. Many who did not file used other programs, like sick leave. Similarly, Morse, Dillon, Warren, Levenstein, and Warren (1998) conducted a random-digit-dial survey in Connecticut to determine the period prevalence of work-related upper extremity musculoskeletal disorders. They also asked respondents whether they had filed for workers' compensation. Of cases in which a medical provider had identified the condition as work-related, only 21 percent of respondents said that they had filed for workers' compensation benefits. Finally, a study by Lakdawalla and Reville (2001) based on the National Longitudinal Survey of Youth indicates that only 55 percent of reported occupational injuries resulted in workers' compensation claims.

Not even fatal injuries are completely recorded by the systems designed to capture them. Stout and Bell (1991) review studies of the completeness of fatal injury reporting, using OSHA fatality reports, workers' compensation records, death certificates, medical examiner records, and state health department records. Between 40 and 70 percent of occupational fatalities identified by at least one source were captured by workers' compensation, and between 21 and 42 percent were captured by OSHA fatality reports. (Of course, death certificates and medical examiner records cannot be used for surveillance of non-fatal injuries.)

Studies have shown not only that reporting is incomplete but also that it is affected by the incentives faced by those who report. Ruser and Smith (1988) studied the impact of OSHA's records-inspection policy of the early 1980s. They estimated that this policy, by which OSHA did not inspect workplaces with low reported injury rates, led to a decline in reporting of 5 to 14 percent. Boden and Gold (1984) examined reporting of coal dust exposures by mines subject to regulatory penalties based on samples collected by the mine operators. They found evidence of widespread underreporting, with most mines providing unrealistically low samples at least 15-30 percent of the time. Recently, Boden and Ruser (2004) have shown that changes in workers' compensation statutes in the 1990s to make claim filing more difficult led to a 10 percent reduction in 1997 reports of days-away-from-work injury rates in the changing states.

In some settings, where multiple company-specific surveillance sources are available, combining these sources (occupational clinic records, short-term disability insurance, medical insurance, and so on) with workers' compensation and BLS data can provide a more comprehensive census of workplace injuries. However, such sources are only available at the largest firms, and only a few of these firms keep the information required to allow this. Moreover, the data are private and formatted inconsistently. Another potential data source that could improve measures of the incidence of occupational injuries and illnesses is state occupational disease surveillance programs. These programs can provide important supplemental information, and they have been used to measure workers' compensation filing behavior (Biddle et al. 1998, Biddle and Roberts 2001). They almost certainly provide much better information on chronic conditions than either the BLS SOII or workers' compensation claim data. Still, their coverage is limited in scope and varies considerably from state to state. Other reporting systems, like the OSHA Integrated Management Information System, only contain aggregate information on the number of injuries in a workplace by category. For this reason, they cannot be used in a capture-recapture analysis. As a consequence, improving how we use the BLS SOII and workers' compensation data offer the greatest potential for improving our national system of non-fatal injury surveillance.

Recent studies have attempted to generate better measures of the frequency of occupational injuries and illnesses using capture-recapture methods rates (Cormack, Chang, and Smith 2000; Maizlish, Rudolph, Dervin, and Sankaranarayan 1995; Morse, Dillon, Warren, Hall, and Hovey 2001; Alho 1990). Capture-recapture methods originally were developed to estimate the population size of animals living in the wild. Sequential independent samples of the populations of interest are captured, tagged, and released. After the first sample, researchers can calculate the total population size from the number of tagged animals recaptured and the number uniquely captured in the initial and each subsequent sample. Observing the

formal equivalence of the problem of estimating animal populations to that of estimating injury or disease frequency from incomplete population-based registries, epidemiologists have applied capture-recapture methods to a wide variety of conditions. We use these methods to obtain better estimates of non-fatal occupational injury frequencies.

In the next section, we describe the two data sources on workplace injuries and illnesses that we rely on to determine the degree of completeness of current reporting: state workers' compensation data and the annual Bureau of Labor Statistics survey.

DATA AND METHODS

Data

State workers' compensation data

We have obtained electronic data for 1998-2001on reported workers' compensation cases from 6 states: Minnesota, New Mexico, Oregon, Washington, Wisconsin, and West Virginia. For each case, we have the injured worker's name, the number of lost days paid for each claim, the worker's age, gender, length of service (for 5 of the states), the part of body that was injure, the nature of injury (for example, sprain, fracture, hearing loss, and so on), the type of ownership of the firm (public/private), the date of injury, either the state identification number or the federal employer identification number (EIN), the 4-dight SIC, and the employer's name and address.

Washington State and New Mexico collect data on all cases involving medical care or lost time. Other states in this study collect data only on cases exceeding the waiting period (the number of days off work that must be exceeded for a worker to collect workers' compensation income benefits (Table 1).

Table 1. Information about State Workers' Compensation Data

State	Waiting period (days)	Range of cases for which data are available
Washington	3	All cases
West Virginia	3	All cases
New Mexico	7	All cases
Minnesota	3	Lost-time only
Oregon	3	Lost-time only
Wisconsin	3	Lost-time only

BLS Survey of Injuries and Illnesses (SOII) Data

Our second source of injury and illness data is the BLS SOII, an annual, nationally representative survey of 165,000 private industry establishments conducted by the U.S. Bureau of Labor Statistics. The

injury data are based on logs that establishments maintain as required by the Occupational Safety and Health Act of 1970.[2] Surveyed establishments report annual counts of all recordable injuries and of three specific types of nonfatal injuries, as well as the total annual hours worked by all employees, the number of employees, and the establishment's industry. The injury and hours data are used to estimate injury rates, expressed as the number of injury cases per 100 full-time worker years.

In order of increasing severity, the three specific types of reported injuries are: non-lost-workday, restricted-workday-only, and days-away-from-work.[3] Workers with non-lost-workday injuries return to all of their duties by the day following the injury. A restricted-workday-only case involves some restriction on the amount of work that workers can perform on the day or days following the injury, but where they return to work by the day following the injury. In contrast, when workers sustain days-away-from-work injuries, they do not return to work on the day or days following the injury (although some of the lost workdays may be days at work but with restrictions).

Beginning in 1992, the BLS has required surveyed establishments to report detailed case-level data for days-away-from-work injuries. For workplaces with more than 20 expected days-away-from-work injuries, the BLS does not request detailed data on all injuries. Instead, it asks the employer to report on injuries that occurred over a specified period that is shorter than one year.

The BLS SOII data include for each reported injury: name of injured worker, gender, age, length of service at the establishment (not required), date of injury, days of lost work, days of restricted work, nature of injury (sprain, fracture, amputation, etc.), part of body affected, and type of injury (fall, overexertion, etc.)

Methods

Creating comparable samples

Lost time cases

If a case type is captured by one source but not another, then it will always appear as unmatched. For this reason, capture-recapture methods should only analyze case types common to all sources. Yet, workers' compensation systems and the BLS survey differ in their definitions of reportable injuries. The BLS survey captures data at the individual level only for days-away-from-work cases – where the worker has not returned to work by the day after the injury. In contrast, workers' compensation lost-time injuries typically must pass the waiting-period threshold, which is either 3 or 7 days in the states in this study. Workers' compensation medical-only claims must have medical-care benefit payments, but they may involve no days lost from work.

In the three state workers' compensation systems that capture both medical-only and lost-time claims, (Washington, New Mexico, and West Virginia), the range of cases covered by both BLS and workers' compensation is similar. Still, workers' compensation data can include cases lacking days away from work. For workers' compensation medical-only cases, information is not available on the number of days lost, so it is not possible to determine whether medical-only workers' compensation claims meet the

[2] Establishments with 10 or fewer employees and those in certain low hazard industries are not required routinely to maintain injury logs except when they are prenotified to do so because they have been selected for participation in the BLS survey.

[3] Before 1992, the survey collected fatality counts. Beginning in 1992, these counts have been collected in a separate census, the Census of Fatal Occupational Injuries. There are very few fatalities relative to the total number of non-fatal injuries, about 6,200 fatalities compared to 6.1 million non-fatal injuries in 1997, so there is little difference between the survey estimates with and without fatalities.

BLS criteria for days-away-from-work cases. For these three states (and the others reporting only lost-time cases), we therefore include only lost-time claims in the sample to be matched.

Given our workers' compensation samples, matching BLS data to workers' compensation claims involves determining which BLS cases would be considered lost-time claims in workers' compensation. For each state, we have developed a program that culls out cases from the SOII data that do not meet the state's rules for lost-time cases. After matching, we apply this rule to unmatched BLS cases to remove those not in the WC reporting frame.[4]

Establishment Samples

A second challenge is the fact that the BLS survey is on a sample of establishments within each state. Workers' compensation reporting systems do not identify specific establishments for reporting purposes, but typically report by firm. So we can identify firms in the workers' compensation data that report to BLS, but we cannot identify reporting and non-reporting establishments within those firms. For single-establishment firms, this is not a problem. However, for multi-establishment firms (for example grocery chains, pharmacy chains, fast food chains), this differential reporting presents a major challenge: In a firm where only a subset of establishments were asked to report, we don't know whether an injury went unreported to BLS because a reporting establishment failed to report it or because the establishment was not asked to report injuries. This means that an estimate of underreporting in workers' compensation could be much higher than actual underreporting. Our approach to this problem uses information from the BLS longitudinal database (LDB) of establishments in the U.S. The LDB provides both employer identifiers (Federal EIN and state unemployment compensation identifier) and unique establishment identifiers. It also provides quarterly employment information by establishment. We used this establishment information to derive the proportion of employment in each firm within establishments reporting to the BLS. Under the assumption that the injury rate per 100 workers is the same within each firm and that the match rate is the same for all establishments within a firm, let:

W	WC reported injuries from all establishments in the firm of interest
W_1	WC reported injuries from all establishments reporting to the BLS in the firm of interest.
L_1	Employment in all establishments reporting to the BLS in the firm of interest
L_2	Employment in all establishments not reporting to the BLS in the firm of interest

We assume that:

$$W_1 = W^* L_1/ (L_1 + L_2)$$

Also,

$W_1 = B_1^+W_1 + B_1^-W_1$, where

$B_1^+W_1$ WC reported injuries in reporting establishments also reported to BLS

$B_1^-W_1$ WC reported injuries in reporting establishments that are not reported to BLS

Then

$$B_1^-W_1 = W^* L_1/ (L_1 + L_2) - B_1^+W_1$$

[4] We considered applying the rule (which will not be perfectly accurate) to all BLS cases, thus potentially discarding BLS cases that match workers' compensation claims, but we instead chose the reported decision rule because it would bias our results toward higher, rather than lower, match rates.

We know all the variables on the right hand side, so we can calculate $B_1^- W_1$. This is the expected number of injuries reported to WC but not to BLS in establishments reporting to BLS under the assumption that selection of reporting establishments within firms is independent of the latent injury rates of these establishments. This quantity is equal to the number of injuries reported to workers' compensation but not BLS only if $L_2 = 0$, that is only for employers all of whose establishments are sampled by BLS (which, by definition, includes sampled single-establishment employers). For other employers, we weight all injuries in the set of unmatched WC injuries for each employer so that the weighted sum within the employer equals $W^* L_1/ (L_1 + L_2) - B_1^+ W_1^+$ (before BLS sampling-based weighting) unlinked workers' compensation injuries. To do this, we multiply the weight of each unmatched WC observation by $[W^* L_1/ (L_1 + L_2) - B_1^+ W_1] / [W - B_1^+ W_1]$.

Industries Examined

Establishments in some industries do not report directly to BLS for its annual survey, so their reporting patterns may differ from those of industries reporting direct to the BLS. This suggests separate analysis of these data. For this reason, we have excluded mining and railroad transportation from this study. Also, state and local employees are not surveyed for the BLS national estimates and some states collect these data while other do not. We exclude these workers as well. The water transportation industry is covered by its own workers' compensation system (as are railroad workers), so this industry is not included in this study. Finally, because of spotty coverage by workers' compensation programs, we also exclude membership organizations. These exclusions reduce the coverage of the analysis in this paper, but, for the industries analyzed, should lead to better estimates.

Matching

We then match these claims to the BLS-reported injuries using data elements common to both sources, including employer identifier; employer name, employer address, employer zip code, city, worker's first initial, worker's last name[5], gender, and date of birth or age at injury. We first do a deterministic match, considering two injuries to be matched if they match on eight of these items. We chose to match on eight factors because we looked at the matched cases and found that this rule always produced items that appeared properly matched.

For remaining unmatched cases, we apply probabilistic record linkage, which is widely used by, among others, the U.S. Bureau of the Census. It is also the most common matching method used in linking medical records from different sources. Fellegi and Sunter (1969) and Copas and Hilton (1990), among others, present a formal theory for probabilistic record linkage, which operates by assigning weights to information fields in potentially matched pairs and then developing a combined weight for each pair. A large weight indicates a high probability that both of the pair's records refer to the same event. For example, a pair matched by name in which both names were Allen Kurzweil would receive a higher agreement weight than if the both names were John Smith. Jaro (1995) develops weights as follows:

agreement weight = $\log_2$[sensitivity/(1-specificity)]

disagreement weight = $\log_2$[(1-sensitivity)/specificity]

Sensitivity is defined as the probability that the field matches given that the pair refers to the same event. Specificity is defined as the probability that the field does not match given that the pair does not refer to the same event. To be a useful field, sensitivity must be greater than (1-specificity). If this is the

[5] We allow flexibility in matching the last name, using the soundex code. Also, first initial is used because the BLS only codes this, not the full first name.

case, agreement weights are positive and disagreement weights are negative. If several fields were used in the match, each would have its own agreement and disagreement weights. A composite weight for a pair is the sum of its agreement and disagreement weights. Note that agreement weights will always be positive, and disagreement weights will always be negative. Sensitivity and specificity can be calculated iteratively from the lists to be matched (Jaro 1995).

Probabilistic record linkage establishes two cut-off points. If a weight is above the high cutoff point, the pair is considered matched. If it is below the low cutoff point, the pair is considered unmatched. Pairs between the two are then hand-matched (on the assumption that hand-matching is the most accurate method available). In our situation, the hand matching was done by two coders and compared. Where the coders differ (across the states, concordance measured by Cohen's kappa ranged between 0.79 and 0.84), we apply a decision rule (available on request) to decide whether a pair was matched. Among the 6 states, probabilistic linking produced about 5-10 percent of all linked cases.

Capture-recapture

Capture-recapture methods are used in epidemiology to estimate disease incidence or prevalence from multiple, overlapping, but incomplete sources. Typically, capture-recapture analyses identify all unique cases recorded by at least one source and then use log-linear or logistic models to estimate the number of cases unidentified by any source. The estimate of the missing cases typically is based on the assumption that the sources are independent samples of the target population. The idea behind these models is simple. Table 3 shows hypothetical results of injury surveillance using two data sources, where X is the unknown number of missed cases. If capture by Source 1 and Source 2 is independent, then Source 1 will capture the same fraction of cases among those captured by Source 2 and those not captured by Source 2. That is,

$$N_1 / (N_1 + X) = N_3 / (N_3 + N_2) \text{ or } X = N_1 N_2 / N_3 \quad (1)$$

This is the maximum-likelihood estimator of X under the assumptions of independence of ascertainment, uniform probability of ascertainment within both sources, and a closed population (Hook and Regal 1995). If the sources are (positively) dependent, then the probability of being in both sources is greater than the product of the probabilities of being only in one source or the other. Source 1 will capture a greater proportion of cases captured by Source 2 than those not captured by Source 2. That is,

$$N_1 / X < N_3 / N_2 \text{ or } X > N_1 N_2 / N_3 \quad (1a)$$

Therefore, Equation (1) will underestimate X. Similarly, if the sources are negatively dependent, the probability of being in both is less than the product of the probabilities of being in either.

Table 2. Two-source Capture-Recapture Model

Source 1		Source 2	
	No	Yes	Total
No	X	N_2	$X + N_2$
Yes	N_1	N_3	$N_1 + N_3 = S_1$
Total	$X + N_1$	$N_2 + N_3 = S_2$	$X + N_1 + N_2 + N_3 = S$

Although originally developed in wildlife biology and demography, capture-recapture methods have been used to generate improved estimates of drug use, homelessness, infectious diseases, diabetes, to name a few

(International Working Group for Disease Monitoring and Forecasting 1995). In the occupational health and safety arena, they have been used to estimate occupational fatality rates (Cormack, Chang, and Smith 2000), carpal tunnel incidence (Maizlish, Rudolph, Dervin, and Sankaranarayan 1995), the incidence of work-related musculoskeletal disorders (Morse, Dillon, Warren, Hall, and Hovey 2001), and the period prevalence of occupational diseases (Alho 1990). They have not yet been used to estimate overall nonfatal occupational injury and illness rates.

To generate unbiased estimates of the total population, capture-recapture studies often make several assumptions in addition to the non-dependence of sources. These include: that the population remains closed (constant) over the observed period and that each individual has the same probability of capture. Although the assumption of a closed population appears reasonable for non-fatal occupational injury surveillance, we expect that capture probabilities will vary by employer, worker, and injury characteristics. For this reason, we will estimate the capture probability conditional on these characteristics. This allows us to develop a better measure of injury incidence.

We account for observed heterogeneity of capture probabilities using logistic regression. Several articles describe the favorable characteristics of logistic regression models in this context (e.g., Ahlo 1990 and Tilling and Sterne 1999). We first use the multinomial logistic model to estimate the probabilities of being observed by workers' compensation alone, BLS alone, or both, conditional on being observed. All statistical analysis use BLS state weights to account for establishment sampling and for sampling of injuries within establishments.

There are three probabilities to be estimated, p_1, p_2, and p_3, equal respectively to $N_1/(N_1+N_2+N_3)$, $N_2/(N_1+N_2+N_3)$, and $N_3/(N_1+N_2+N_3)$ in Table 3. The multinomial logistic model for two-source capture-recapture estimation without covariates is then:

$$\Pr(Y_i = j) = p_{ij} = e^{\beta_{0j}} / \sum_{k=1}^{3} e^{\beta_{0j}},\ j = 1,2,3 \qquad (2)$$

where Y_i indicates cell membership for the i^{th} observed subject. To identify the model, one of the parameters is set to zero, and the model is then estimated. The parameter estimates then can be used to generate an expected probability for each captured individual to be captured by Source 1, Source 2, or both sources. This also leads to an estimate for the probability of being missed by both sources.

To account for heterogeneity of capture probabilities according to individual, injury, and employer characteristics, the multinomial logistic model can be generalized to include a vector of covariates:

$$\Pr(Y_i = j) = p_{ij} = e^{\beta_{0j}+\beta_j' x_i} / \sum_{k=1}^{3} e^{\beta_{0j}+\beta_k' x_i},\ j = 1,2,3 \qquad (3)$$

The parameter estimates can be estimated through maximum likelihood and then be used to generate an expected probability for each individual to be reported by Source 1, Source 2, or both sources contingent on being reported by at least one. Given the assumption of independent capture by both sources for each individual i, we can estimate p_{i0}, the probability of being missed for person i (Tilling and Sterne 1999).

The total population size S is then estimated from the observed sample as:

$$S = \sum_{i \in N_j}^{j=1,3} (1/(1-p_{i0}))\ ) \qquad (4)$$

We used multivariate multinomial logistic regression to generate estimates of p_1, p_2, and p_3, using as independent variables: employment size categories, average wages per hour worked in the establishment, 1-

digit SIC, age, sex, tenure, part of body, and nature of injury.[6] We developed separate estimates for each state.

RESULTS

Table 1 shows the estimates of reporting completeness for cases meeting the workers' compensation threshold for lost time injuries. Because these estimates assume source independence, they probably overestimate actual completeness of ascertainment. Several things are clear from this table. First, there is substantial variation in completeness of both BLS reporting and workers' compensation reporting among states. That workers' compensation reporting completeness varies a great deal should not be very surprising, as workers' compensation systems differ substantially from state to state. However, the BLS annual survey is a national program with a single set of reporting guidelines, so one might not expect considerable interstate variation. Second, looking across states, there is a surprisingly low correlation between workers' compensation and BLS reporting rates. States that have little underreporting to their state workers' compensation system may have moderate to high underreporting to the BLS survey and vice versa. Also, combining injuries from the two sources improves the estimate of overall ascertainment to at least 86% in all states.

Table 3. Estimates of Reporting Completeness for 6 States:
Lost Work Longer than the Workers' Compensation Waiting Period

Percent reported to:	WA	WV	OR	WI	NM*	MN
Workers' Compensation	94%	91%	77%	75%	67%	64%
Bureau of Labor Statistics	51%	75%	55%	65%	50%	71%
Either – Adjusted Estimate	96%	97%	88%	92%	84%	88%
Crude Estimate	97%	98%	90%	91%	82%	90%

Note: Calculations assume source independence and are adjusted for age, sex, job tenure, industry, establishment size, body part injured, and nature of injury. 95% confidence intervals extend less than 1.5 percentage points from the point estimate for all states but New Mexico, where they extend about 3 percentage points.
*New Mexico has a 7-day waiting period. The other states have 3-day waiting periods.

Sensitivity to source dependence

Our estimates depend on the assumption of source independence. Yet we expect some degree of positive source dependence between the two sources we propose to use. In some workplaces, the same individual may be responsible for filling out the BLS survey and for filing workers' compensation claims. Even if this is not the case, an employer may believe that an injury that is not considered compensable by the workers' compensation system should not be reported to the BLS. In states with restrictive workers' compensation compensability requirements, relatively few injuries may be reported to the workers' compensation system (Boden and Ruser 2004, Messiou and Zaidman 2005). But, one study has found that reporting to the BLS is also suppressed. The decline in BLS reporting after states pass such laws (Boden

[6] Because several injury coding methods were used, we recoded part of body and nature of injury into compatible broad categories.

and Ruser 2004) is consistent with this influence of workers' compensation laws and regulations on both workers' compensation claim rates and BLS injury reporting.

We cannot estimate the degree of dependence of the two sources without additional assumptions or a third data source. Brenner (1994, Figure 1) demonstrates that variation in positive source dependence typically produces differences in estimates of incidence that are not very large. Moreover, all estimates are biased in the same direction, so the difference in bias between states will be smaller than the bias in any state.

Given positive source dependence and appropriate correction for source mismatch, we know at least three things about the estimates we derive: (1) that they are better than the incidence derived from each source by itself, (2) that they are better than estimates derived from adding the number of cases reported by each source alone to the number reported by both, and (3) that they are a lower-bound estimate of the true number of cases. Thus, this study can achieve substantial improvement in incidence estimates, despite the fact that the estimates will be biased.

To investigate the sensitivity of estimates to positive source dependence, we have derived population size estimates under a range of dependence scenarios. With source independence, the odds that workers' compensation ascertains an injury are independent of whether the injury is reported to BLS. Thus, the odds ratio (the ratio of the odds of reporting to workers' compensation conditional on reporting to BLS to the odds of reporting to workers' compensation conditional on not reporting to BLS) is one. We can re-estimate the number of unreported injuries under alternate assumptions about the value of the odds ratio (OR). Table 5 shows the results of this sensitivity analysis for OR = 1.0, 3.0, and 5.0. We can see from this table that source dependence reduces the estimated coverage of the two sources of injury data.

Table 4. Estimates of Reporting Completeness for 6 States: Impact of Assumptions about Source Dependence

Percent reported to:	Odds ratio*	WA	WV	OR	WI	NM†	MN
Workers' Compensation	OR=1	94%	91%	77%	75%	67%	64%
	OR=3	88%	88%	67%	67%	53%	57%
	OR=5	83%	85%	61%	61%	45%	52%
Bureau of Labor Statistics	OR=1	51%	75%	55%	65%	50%	71%
	OR=3	50%	73%	48%	58%	42%	61%
	OR=5	49%	71%	44%	53%	37%	55%
Either	OR=1	96%	97%	88%	92%	84%	88%
	OR=3	90%	93%	75%	81%	67%	75%
	OR=5	86%	91%	68%	74%	57%	68%

*The ratio of the odds that an injury is reported to workers' compensation if it is reported to BLS to the odds that an injury is reported to workers' compensation if it is not reported to BLS. This ratio equals one for source independence and is greater than one for positive source dependence.

Factors affecting reporting

The coefficients of the multinomial logistic regression that estimates the probability of an injury's being in both sources, just the SOII, or just the state workers' compensation data are not directly interpretable, so we take the approach of comparing the predicted values of reporting completeness for

alternate employer, injury, and worker characteristics for each state. We begin with a base case: a back strain that occurred in a manufacturing establishment with 100-250 workers to a male worker aged 35-44. We then predict the effect of changing covariates one at a time.

There is little state-to-state consistency in the impact on ascertainment of many of the factors we examined. However, some alterations from the base case produce changes of more than 5 percent in the probability of ascertainment for more than half the states.

Differential ascertainment for specific characteristics, occurred mostly in the BLS survey. In all 6 states, the smallest firms, those with 11-49 employees reported at least 5 percent more than did the mid-sized firms in the base case. Wholesale and retail firms had lower ascertainment than manufacturing firms in 5 of 6 states. In addition, upper extremity cumulative trauma disorders had lower rates of ascertainment in 5 of 6 states. In 3 of the states, ascertainment was much lower (38% to 70% lower). Chronic occupational disease was reported so infrequently that ascertainment could not be estimated in 3 of 6 states. Even where estimates were made, we do not think they are reliable.

The only factor with a greater than 5 percent impact on workers' compensation injury ascertainment in more than half the states is age. In the age 55+ group, injury ascertainment is at least 5% higher in 4 of the 6 states.

SUMMARY OF SPECIFIC AIMS OF THE GRANT AND SUCCESS IN MEETING THEM

(1) Obtain and prepare workers' compensation databases from participating states. For 1993-2000 injuries, we will match state workers' compensation data to BLS injury data by employer and injury.
COMPLETED

(2) Measure overlap between workers' compensation and BLS data and develop an algorithm for matching lost-time workers' compensation injuries to BLS injuries. We will use 1998-2000 data from Washington, New Mexico, and West Virginia, where all workers' compensation claims (including medical-only claims) are reported and from Texas, which reports all claims with at least one day away from work. Then, we will match BLS and workers' compensation data for all states for 1998-2000. We will determine whether matching would be feasible for 1993-1997 injuries, for which the BLS has discarded injured workers' names.
SUBSTANTIALLY COMPLETED. Texas data were too poor to use, but other states' data were successfully matched. Matching with a high degree of confidence was only possible for 1998-2001 (one year added when it became available).

(3) Compare workers' compensation claims with BLS-reported injuries in each participating state. We will develop a lower-bound measure of annual injury incidence using capture-recapture methods. Using logistic regression, we will determine the factors associated with differential capture probabilities of the workers' compensation and BLS sources for employer, worker, and injury characteristics.
COMPLETED.

(4) Do a sensitivity analysis. simulate the effect of a range of mismatch rates; describe the effect of a range of source dependence on the estimates of total injury incidence.
COMPLETED.

(5) Determine which specific types of employers, workers, and injuries appear most likely to be missed by the workers' compensation and/or BLS systems.
COMPLETED.

(6) Compare results across states; look for similarities and differences across states in overall capture rates and in the variation of capture rates across employer, worker, and injury characteristics.
COMPLETED.

(7) Determine the extent to which the results of this study may help to improve estimates of nonfatal occupational injury and illness incidence in the U.S. as a whole.
PARTIALLY COMPLETED. The variation in state reporting is substantial, and we do not yet understand the reasons for this variation. We are working with the BLS to see if we can design a study to help pinpoint the problems with BLS data so that its completeness can be improved. We have also determined that using state workers' compensation and BLS data together can substantially improve state estimates.

PUBLICATIONS

No papers have been published yet. One has been sent back to the journal after responding to reviewers' comments: Capture-Recapture Estimates of Nonfatal Workplace Injury Rates. I will inform NIOSH when this is published.

Inclusion of Gender and Minority Subjects

All workers with reported injuries are included in this study, so male and female workers and minority workers are included if they have reported injuries.

Inclusion of Children

All workers with reported injuries are included in this study, so workers younger than 18 are included if they have reported injuries.

Materials Available for other Investigators

The BLS only allows its data to be used on-site with its written permission. The investigators lack materials to share with other investigators.