Adjusting for the Healthy Worker Selection Effect in Cross-Sectional Studies

LAURA PUNNETT


Background. In a cross-sectional study of musculoskeletal disorders, women employed in highly repetitive manual work (garment assembly) were found to have approximately double the risk observed in a population with more varied tasks (hospital work). It was suspected that this estimate might be biased if garment workers with musculoskeletal pain were more likely than others to leave employment.

Methods. Retrospective information on date of first onset of symptoms and years employed to date of pain onset, or to survey date (whichever was earlier), was used to calculate age and calendar period-specific rates of onset, conditional on remaining employed until the survey.

Results. These rates, and the relative risk for garment work, increased over the 20-year period preceding the year of the survey. The trend was not explained by age or length of employment, or by any known changes in work demands that might have caused a true increase in incidence density.

Conclusions. In the absence of longitudinal cohort data, alternative explanations for these results cannot be excluded. However, with specified assumptions, the most plausible appears to be a healthy worker selection effect acting differentially between high and low exposure groups. This effect would have caused the smallest bias in the prevalence in the year immediately before the survey, and a better estimate of the true relative risk would be approximately five. Where date of onset has been obtained, this method may be used in other cross-sectional studies to estimate and reduce the magnitude of selection bias in a survivor population, if longitudinal data cannot be collected.

Keywords: cross-sectional studies, epidemiological methods, healthy worker effect, occupational disease, selection bias
the resulting bias in the estimation of exposure effect can only be eliminated by a longitudinal study of the full cohort. In such a longitudinal study, the causal effect of exposure can be measured in terms of the ratio of age-calendar period-specific rates of disease onset between exposed and unexposed subjects. In a cross-sectional study, it is not possible to calculate these rates directly. However, the data in this study permitted the incidence density rates to be estimated empirically from the age-calendar period-specific rates of disease onset, conditional on being an active worker at the time of the survey. Such rates are here labelled 'conditional incidence rates'.

The goal of this paper is to evaluate whether cross-sectional data on active workers can, given certain reasonable assumptions, be used to estimate the exposure effects that would have been observed in a longitudinal study of both complete cohorts. Although the example presented here concerns the effects of repetitive manual work on musculoskeletal pain, these methods could be applied to cross-sectional studies on any morbid outcome for which a relatively well-defined date of onset can be identified by the subjects or from medical records. Such outcomes might include asthma, acute change in forced expiratory volume (FEV1), sensitization reactions, dermatitis, angina pectoris, and peripheral neuropathy such as that induced by lead exposure.

MATERIALS AND METHODS
Population and Data Collection
The study population consisted of 162 women garment workers who completed a questionnaire in the period from late December 1981 to early January 1982. Medical, work and symptom histories were obtained from all participants. The jobs of the garment workers were analysed and found to involve high exposure to repetitive manual work, performed under a piece-rate wage system, with postural stress and static muscular loading.

The comparison population consisted of 73 women hospital workers, none of whom performed typing or other repetitive manual tasks for ≥4 hours a day. The hospital workers completed the same questionnaire. Job descriptions and work observations were used to attempt to ensure that only hospital employees with low ergonomic exposures were included in the comparison group.

The key items on the questionnaire were: 1) whether the subject had ever experienced pain, numbness or tingling of the back, neck, shoulder, elbow, wrist or hand that lasted for most days of one month or more; 2) whether she had experienced those symptoms within the last year; 3) the calendar year of first onset of pain; and 4) the date of first employment in the current industry. Prevalent pain was defined as that suffered for most days of one month or more, within the last year, that was not associated with prior injury to the same site, and that began after first employment in garment manufacture or hospital work, respectively.

Four garment and four hospital workers with prevalent pain did not report date of first onset and were excluded from this analysis. This left 158 garment workers and 69 hospital workers in the study and comparison populations, respectively. Additional information on the study participants and survey methods has been reported.

Data Analysis
All analyses were performed using the SAS and BMDP software packages for mainframe computers. For purposes of analysis, the date of the survey was defined as 1 January 1982. Calendar year of onset, t, was denoted by the number of years preceding the study date, t = 0 to 20, where t = 0 at the time of the survey, t = 1 in the year preceding the survey (1981), t = 2 in the previous year (1980), etc.

The annual incidence density of pain in each industry was approximated by the probability of developing pain in calendar year t, conditional on being employed in year t, being at risk to develop pain for the first time in that year, and remaining employed until the date of the survey. This conditional probability was estimated by the ratio of the number of women reporting first onset of pain in a given calendar year to the number of women employed in that year who had not developed prevalent pain in previous years. The denominator for each year t, in each workplace, was computed by summing the number of workers from the cross-sectional population who had been employed in that year and at risk for new pain, i.e. with no incident reported for any previous year. The corresponding numerator was computed by summing the number of workers employed and at risk who reported a first incident in that year.

The term, 'conditional incidence density', is used here to represent this exposure-specific yearly conditional probability, as estimated from the cross-sectional data. The term, 'relative risk', is used here to refer to the ratio of the conditional probabilities of onset in year t in the garment workers as compared to the hospital workers. Each conditional probability is bounded by 0 and 1, unlike a true incidence density. However, since in this study the maximum conditional probability was 0.16, these conditional probabilities serve as adequate estimations of the true incidence rates among the populations available for study.
The exposure-specific, conditional incidence densities and relative risks for each year were also estimated after stratification by age and by length of employment. The numerators and denominators were computed in the same manner, first, after dividing the number of workers at risk in each year into four strata defined by age in that year, and second, after dividing the workers at risk in each year into three strata defined by seniority in that year. Tests of linear trend in incidence density over time were calculated.

The logit of the annual conditional incidence rates was modelled using proportional hazards analysis, fitted by the Cox regression routine in BMDP. The outcome was posited as a function of type of workplace ('X'), year of onset ('t'), age stratum ('Age'), length of employment stratum ('Yrs'), and an interaction term for exposure and year of onset ('X*t'). The following multivariate models were fit:

A. \[
\ln(p_t/1-p_t) = B_0 + B_1*X + B_2*t + B_3*X*t
\]
B. \[
\ln(p_t/1-p_t) = B_0 + B_1*X + B_2*t + B_3*X**t + B_4*Age_t
\]
C. \[
\ln(p_t/1-p_t) = B_0 + B_1*X + B_2*t + B_3*X**t + B_4*Yrs_t
\]

where \( p_t \) = the conditional probability of developing persistent pain in year \( t \)
\( X \) = 1 for garment work, 0 for hospital work
\( t \) = the number of years prior to the date of the survey, \( t = 0 \) to 20
\( Age_t \) = 1 to 4 for age stratum in year \( t \): <30 years old; 40-49 years old; ≥50 years old
\( Yrs_t \) = 1 to 3 for seniority stratum in year \( t \): 1-5 years; 6-10 years, ≥11 years

The regression coefficients in a proportional hazards model, as in logistic regression analysis, are interpreted as the log odds ratios. The contribution of adding a new term to a model in a nested series is evaluated by the likelihood ratio test.

In each of the models listed above, \( B_1 \) is an estimate of the relative risk for exposure to garment assembly work in year \( t = 0 \), i.e. at the moment that the survey took place. The coefficient, \( B_2 \), is an estimate of the linear change in log incidence density among the hospital workers with increasing proximity in calendar time to the year of the survey. The coefficient, \( B_3 \), or 'interaction term', is an estimate of the linear change in the log relative risk with increasing proximity to the survey date. Therefore, the odds ratio for garment workers versus hospital workers in year \( t \) is \( \text{OR} = \exp(B_1 + t*B_2) \). The odds ratio for the effect on incidence density of year \( t \), compared with year \( t - 1 \), is estimated by \( \text{OR} = \exp(t*(B_2 + B_3)) \) in the garment workers and \( \text{OR} = \exp(t*B_3) \) in the hospital workers. Models B and C adjust for the potential confounding effects of age and length of employment on conditional incidence density, respectively.

RESULTS
The crude prevalence ratio for pain at one or more locations in the year before the survey was 1.9 (95% CI: 1.2-2.9) among the garment workers compared with the hospital workers. (This point estimate was revised upwards after controlling for several covariates.) The prevalence odds ratio, which is a better estimate of the true incidence rate ratio for non-rare diseases, was 2.6 (95% CI: 1.4-4.9).

All of the cases of prevalent pain reported first onset within the 20 years prior to the study date, and there were virtually no cases with onset more than 10 years prior to the study date. The remainder of the analysis was therefore restricted to cases occurring in the decade (1972-1981) immediately preceding the date of the survey.

In the garment shop, the yearly conditional incidence density was 0.16 in the year immediately preceding the survey and decreased with distance in calendar time from the date of the survey (Figure 1). A test of linear trend in incidence density over 10 years was highly significant (\( \chi^2 = 33.8 \) on 1 d.f., \( P < 0.0001 \)). The yearly incidence rates in the hospital workers, which were based on sparse data, displayed no obvious trend with calendar time, and the test of linear trend was not statistically significant (\( \chi^2 = 0.835 \) on 1 d.f., \( P = 0.18 \)).

The point estimates of the risk ratio and the odds ratio for garment versus hospital work were both approximately 5 in the most recent year (\( t = 1 \)) and fell below unity at 5 years prior to the survey (Table 1). The odds ratio for garment versus hospital work, controlling for year of onset, was also approximately 5 in the proportional hazard model (Table 2). The coefficient for \( B_3 \) was negative and statistically significant, indicating that the relative risk decreased by about 75% in each successive year prior to the survey date. This was a less steep decline than that observed in the crude odds ratio over the entire 10-year period (Table 1), suggesting that the relationship with the log odds might not be truly linear. However, the model fit the crude data fairly well in years \( t = 1, 2, \) and 3, where most of the data points were concentrated.

The calculation of age-stratified rates and the addition of age stratum to the regression model had virtually no effect on either the point estimate or its standard error for any of the first three coefficients.
FIGURE 1  Conditional incidence density (see text) of musculoskeletal pain, by calendar year, among 158 female garment assembly workers and 69 female hospital employees, Massachusetts, USA, 1972–1981

TABLE 1  Conditional incidence densities, rate ratios, and odds ratios of musculoskeletal pain in 158 female garment assembly workers and 69 female hospital employees, by year prior to survey date, 1972–1981

<table>
<thead>
<tr>
<th>Year prior to survey</th>
<th>Garment workers</th>
<th>Hospital workers</th>
<th>Rate ratio</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.161 (18/112)</td>
<td>0.034 (2/59)</td>
<td>4.7</td>
<td>5.5</td>
</tr>
<tr>
<td>2</td>
<td>0.125 (15/120)</td>
<td>0.034 (2/59)</td>
<td>3.7</td>
<td>4.1</td>
</tr>
<tr>
<td>3</td>
<td>0.095 (11/116)</td>
<td>0.041 (2/49)</td>
<td>2.3</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>0.047 (5/107)</td>
<td>0.049 (2/41)</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>0.021 (2/94)</td>
<td>0.088 (3/34)</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>6</td>
<td>0.012 (1/86)</td>
<td>0 (0/28)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>0.024 (2/83)</td>
<td>0.048 (1/21)</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>0.013 (1/76)</td>
<td>0 (0/19)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>9</td>
<td>0.014 (1/71)</td>
<td>0 (0/18)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>0.016 (1/62)</td>
<td>0 (0/16)</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Note: See text for full definition of 'conditional incidence density'.

Frequencies for previous 10 years (1962–1971) not shown because of very sparse data.

TABLE 2  Cox proportional hazard model of conditional incidence density of musculoskeletal pain in 227 women workers as a function of type of industry (X), year of onset in years prior to survey (t), and interaction between industry and year (X*t), 1972–1981

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Beta</th>
<th>Odds ratio</th>
<th>SE(^b) (Beta)</th>
<th>Beta/SE</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry (X)(^c)</td>
<td>1.63</td>
<td>5.12</td>
<td>0.58</td>
<td>2.83</td>
<td>0.005</td>
</tr>
<tr>
<td>Year of onset (t)</td>
<td>−0.11</td>
<td>0.89</td>
<td>0.12</td>
<td>−0.92</td>
<td>0.36</td>
</tr>
<tr>
<td>X(^*)t</td>
<td>−0.28</td>
<td>0.76</td>
<td>0.14</td>
<td>−1.96</td>
<td>0.05</td>
</tr>
<tr>
<td>Intercept</td>
<td>−2.89</td>
<td>–</td>
<td>0.52</td>
<td>−5.53</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Note: See text for full definition.

\(^b\) Standard error.

\(^c\) Coded 1 for garment workers, 0 for hospital workers.
Specifically, the effect of the interaction between calendar year and industry remained the same. Similarly, the calculation of rates stratified on seniority, with the addition of a term for stratum of length of employment to the model, had virtually no effect on either the point estimate of the interaction term, $B_3$, or its standard error.

The two key findings of the analysis, therefore, were:
1) In the garment workers, and less so in the hospital workers, the observed 'conditional incidence density' of musculoskeletal pain increased over the 10-year period prior to the survey.
2) The observed relative risk for the effect of garment assembly work on musculoskeletal pain increased from less than 1 to approximately 5 over the 10-year period prior to the survey.

DISCUSSION
Explanation of Findings
In this re-analysis of a cross-sectional study, exposure-specific incidence densities were estimated in a survivor population, using recalled dates of first employment and first onset of prevalent pain. Prevalent cases were assigned to the year of reported onset, and denominators were constructed from workers at risk for first onset in the same year. These 'conditional incidence rates' increased strongly with calendar time in the more exposed population, with a resulting increase in the estimated relative risk. Neither age nor length of employment explained the associations between workplace and year of onset for relative risk.

Interpretation of these findings requires consideration of the likelihood, either that the observed increase over time truly occurred, or that the true incidence density and relative risk were constant over time, and the observed time trend an artifact of some form of selection or information bias. (The 'true relative risk' is defined as the incidence rate ratio that would have been observed in a longitudinal study of both complete occupational cohorts with no misclassification of outcome. This is the true causal parameter of interest.)

In the absence of the hypothetical cohort study, it cannot be conclusively determined which of these competing hypotheses is true. However, substantive knowledge of the particular populations under study permits consideration of their relative likelihoods. Those explanations that suggest themselves are considered in turn below. Each possible explanation is followed by a statement of how it would have had to operate in order to produce both findings (the increase in incidence density in the garment industry and the associated increase in relative risk with increasing proximity in time to the survey date). The true relative risk that would have been observed in a full cohort study is described, with the underlying incidence rate in the garment cohort illustrated in Figure 2. The plausibility of each argument is then discussed.

A. True Increase (Process Change)
One or more changes may have occurred in work processes or tasks in the garment shop, increasing the physical demands of the work and causing a true
increase in incidence density of musculoskeletal disorders. (If task changes also occurred in the hospital, they would have had less impact on physical work demands, since little or no increase was observed in the incidence density in that population.) A true increase over time in incidence density in the garment workers, and a true increase in relative risk, would have been observed in a full cohort study (Figure 2).

**Plausibility.** This explanation is extremely unlikely. Production processes in the women’s garment industry have changed little in this century. Discussions of process changes with the owner of the garment shop disclosed no recent changes in equipment or work organization that were likely to have increased the incidence of repetitive trauma to the musculoskeletal system or which could account for the sudden onset of symptoms simultaneously in a large number of previously asymptomatic workers.

**C. Non-Persistent Pain**
Prevalent cases of pain included in the cross-sectional survey may have been a mixture of persistent and non-persistent pain, meaning that not all symptoms persisted from the date of onset to the survey date. Cases of non-persistent pain observed in the cross-sectional study, by definition, would have to be of recent onset in order to be observed. If the true incidence density of both types of pain was constant over calendar time in each cohort, then the incidence density in a cross-sectional study would appear to increase with proximity to the date of the survey because of the non-persistent cases. In a full cohort study, if data were collected each year on current symptoms only, no increase in incidence density over time would be observed. The estimated relative risk ratio would be constant over time in both the cross-sectional and the longitudinal studies.

For a case mix of persistent and non-persistent pain to produce the observed data, the true relative risk would have to be greater than unity only for non-persistent but not for persistent pain. Thus, in the cross-sectional study, the risk ratio observed for cases with onset in the most recent year \( t = 1 \) would be based on a combination of persistent and non-persistent pain and would be elevated. However, the relative risk observed for cases with onset, say, five or more years before the survey would be the relative risk for persistent pain only; this would not be elevated. The relative risk observed in the cross-sectional study in the year farthest from the survey date (Figure 2) would be the best estimate of that for persistent pain in the full cohort, assuming that persistent pain was unrelated to remaining in the workforce.

**Plausibility.** In the literature on occupational soft tissue disorders, chronic conditions and long-term sequelae have frequently been reported.\textsuperscript{15-19} In particular, long-term disability has been reported to be elevated among female garment workers, particularly those who had worked under piece-rate wage systems.\textsuperscript{20-22} No studies have suggested an association only for non-persistent pain with repetitive manual tasks and other ergonomic stressors. This appears to be a very unlikely explanation.
D. Healthy Worker Selection Effort
Consider two workers in either cohort, employed in the year prior to the survey, one with soft tissue pain (case) and one without (non-case). The ‘healthy worker selection effect’ is defined to be operating if the probability that the case will leave the workforce prior to the date of study is greater than the probability that the non-case will leave the workforce in the same year.

Suppose that the true incidence density in each cohort was constant over calendar time and that the healthy worker selection effect was stronger in the garment workers than in the hospital workers, although not so strong that only non-diseased garment workers remained in employment. In a cross-sectional study of the ‘survivors’, the incidence density would appear to increase with proximity to the date of the survey, as would the relative risk. The ratio of conditional incidence densities observed in the cross-sectional study would be lower than the ratio of true longitudinal incidence densities, because of cases disproportionately leaving garment assembly work after the incidence of pain. In the full cohort study, the true relative risk and the true incidence density would be observed to remain constant over time. The best estimate of the true relative risk in the cross-sectional study would be that obtained in the year nearest the date of the survey \( t = 1 \) (Figure 2).

Plausibility. It is reasonable to expect that the healthy worker selection effect would operate with greater magnitude among garment workers than among hospital workers. The garment workers experienced pain which was strongly associated with their job features. During interviews, some reported it to be caused or exacerbated by their jobs and others that the pain interfered with job performance; either perception would provide an incentive to stop work. Garment stitchers often have few employment alternatives, whereas some hospital employees may have the option of moving into other job categories (such as administration) if they become physically unable to perform their jobs. A differential healthy worker selection effect, as described here, is both possible and the most plausible explanation of the data.

CONCLUSION
Based on reasonable subject matter assumptions, the most plausible explanation for the observed conditional incidence rates appears to be a healthy worker selection effect that was greater in the highly exposed (garment) than in the less exposed (hospital) workers. Nevertheless, because only cross-sectional data were available, several assumptions were required, as described above and summarized again below.

If differential selection out of employment by health status is the correct interpretation of these data, it has important implications. Many studies of musculoskeletal disorders in workers exposed to ergonomic stressors have been cross-sectional in design. Comparisons between active workers on the basis of current or past exposure may be seriously biased in the direction of underestimating the true adverse effects of exposure when an unknown number of workers are leaving employment because of work-related pain. Analogously, other authors have shown that workers who develop musculoskeletal disorders in jobs with high ergonomic exposures are more likely to transfer to lower-exposure jobs than workers without such disorders. Failure to correct for these selection effects would also bias the relative risk estimates for internal comparisons towards the null value.

In this case, the odds ratio point estimate of 2.6 has been revised upwards substantially on the basis of the analysis presented here. The ratio of conditional incidence densities in year \( t = 1 \), which would have been the least affected by selection bias, was 5.5, and the odds ratio for exposure to garment work from the proportional hazards model, controlling for year of onset and interaction of year with shop (as well as age), was 5.1. Thus, if these results have been interpreted correctly here, the best estimate of the true relative risk that can be obtained from these data is approximately double that estimated in the original study. Furthermore, it appears that the process of developing a work-related musculoskeletal disorder, followed by self-selection out of the workforce, occurred over a period of four years, on average, among these garment workers.

Recommendations for Future Cross-Sectional Studies
Longitudinal studies, with special attention to the follow-up of individuals who have left the workforce for any reason, are always preferable to cross-sectional studies. Only analysis of the full cohort will permit an unbiased estimate of the true magnitude of the health effects of exposure to occupational hazards.

Nevertheless, investigators who only have the resources to conduct cross-sectional studies may be able to refine their estimates of the effect of exposure, as well as the magnitude of the selection bias that resulted from studying only the healthy workers of a cohort. Procedurally, this is a simple matter. A questionnaire is easily expanded with an item pertaining to the date of symptom onset. Conditional disease probabilities then need to be calculated by exposure status in each calendar year; i.e. both cases
(in year of onset) and persons at risk must be assigned to the exposure group to which they belonged in that year. The statistical analysis is then rather straightforward and can be handled by available software packages (e.g. EGRET).

If the results of such analyses show a time trend in relative risk similar to that seen in these data, they could be interpreted as being compatible with a differential healthy worker selection effect under the following conditions: 1) no substantive changes occurred in the relevant occupational exposures over the study period, or, if changes did occur, they have been adequately accounted for in the assignment of historical exposure levels; 2) there is no reason to believe that misclassification of onset date would likely be both systematic and differential, i.e. that the more exposed group would preferentially report later onset dates than actually occurred; 3) all cases of the disorder of interest persist indefinitely after onset, or the true relative risk is at least as likely to be elevated for persistent disorders as for non-persistent ones, or the probability of persisting (i.e. the average duration) is known for both exposure groups and can be accounted for explicitly; and 4) there is no confounding of the observed effect of calendar time by age or seniority.

As a (less efficient) alternative, at a minimum, the effect of selection bias on the observed risk ratio could be reduced by minimizing the time over which the selection effect operates in the data set. The denominator can be restricted to those workers in both shops who had remained pain-free until the recent past (perhaps one year prior to the cross-sectional study), and the numerator can be restricted to those cases that actually occurred within that last year. The restriction would then permit the use of morbidity data on workers still active at the time of the study to approximate more closely the relative risk that might have been observed in a longitudinal study of the two full cohorts.

The approach illustrated here may also be used in a study based on internal comparisons between exposure levels within a single workplace, provided that the exposure level for each individual is known for each calendar year. This will again permit the estimation of a "conditional relative risk", reducing the effect of selection bias on the parameters of interest. However, an alternative approach exists for internal analyses, which is probably simpler to implement and to interpret. Assuming that work histories are adequate to account explicitly for previous job assignments, prevalences may be stratified on whether or not there has been a prior transfer (especially from high to low exposure levels) within a period of time relevant to the latency of the disease under study.

If analyses such as these were carried out more often in occupational cross-sectional studies, the estimates of conditional incidence rates and relative risks for the most recent time interval might represent important supplementary information on the true value of the parameter of aetiologic interest. Additional benefits might accrue in the form of improved knowledge as to the magnitude of healthy worker selection processes in relation to specific combinations of occupational exposures and work-related diseases.

ACKNOWLEDGEMENTS

This research was supported in part by a grant to the Harvard School of Public Health from the Mobil Foundation. Dr Punnett was supported by a National Institute of Environmental Health Sciences National Research Service Award 5T32 ES07069 from the Harvard School of Public Health.

The author is indebted to Dr J M Robins for his contribution to this work. The author also thanks Dr E. Eisen for her valuable comments; D Vincenzo and the staff of the International Ladies' Garment Workers' Union; C Pidcock, M Lyndon, D Plantamura, D Spiegelman and L Pohter for assistance with data collection and analysis; and J Frelich for clerical assistance. An earlier version of this paper was presented at the Vth International Symposium on Epidemiology in Occupational Health, Los Angeles, September 1986.

REFERENCES


(Revised version received December 1995)