CAUSAL ANALYSIS OF BACK BELTS TO PREVENT LOW BACK PAIN

James T. Wassell
Division of Safety Research, National Institute for Occupational Safety and Health,
Centers for Disease Control and Prevention, 1095 Willowdale Road, MS/1811, Morgantown, WV 26505
(email: JWassell@cdc.gov)

Proceedings of the American Statistical Association, Section on Statistics in Epidemiology, Alexandria, VA March 2002. American Statistical Association CD-ROM.

CAUSAL ANALYSIS OF BACK BELTS TO PREVENT LOW BACK PAIN

James T. Wassell
Division of Safety Research, National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, 1095 Willowdale Road, M/S 1811, Morgantown, WV 26505
(email: JWassell@cdc.gov)

KEY WORDS: Injury Epidemiology, Nonrandomized Trial, Propensity Scores, Occupational Health, Prospective Study, Back Belts

ABSTRACT

In a prospective study of the effect of back belts to prevent low back pain, 4,203 material handlers, who did not report a history of previous back pain in a baseline telephone interview, completed a 6-month follow up telephone interview. Belt wearing was self-determined, but influenced by store policy, for which there was poor compliance. Classification tree methods were used to define ten groups of workers with similar propensity to wear back belts. The effect of self-reported belt wearing on self-reported incident low back pain was evaluated within these ten strata. Stratum specific risk differences, odds ratios and logistic regression estimates of the effects of back belts were used to calculate weighted estimates and weighted estimates of variance. There was little evidence to suggest any lack of homogeneity of the effects of belt wearing across strata. Given this propensity stratum adjustment for nonrandom treatment assignment, there was no difference in the incidence of low back pain between those who reported wearing back belts "usually everyday" and those who reported "never" to a question about belt wearing habits.

Introduction

Observational studies to investigate the effectiveness and efficacy of safety equipment in the workplace are limited by various experimental and practical constraints. Randomization of a treatment, in the form of workplace safety apparel or other equipment may not be practical or possible. Further, treatment assignment or self-selection may be influenced by the workplace situation. Given the constraints typical of non-random or quasi-random study situations, there is still a need to evaluate workplace interventions to determine if adverse worker events, such as disabling or fatal injuries, can be better prevented. Statistical models of the underlying process for worker injuries (Wassell, 1998) provide methods to evaluate such workplace interventions. It has long been established that workplace injuries have severe consequences in terms of human suffering and economic losses.

In this paper, some of the concerns of non-random assignment of treatment, specifically self-selection of treatment, which was the case in this study, are addressed using the methods of propensity score adjustment by sub-classification (Rosenbaum and Ruben, 1983). In observational studies, where randomization is not an option for treatment assignment, there is an increased risk that any comparison of the treatment groups may be affected by an imbalance of influential factors. If possible, group randomization (e.g. Kraus et al., 2002), may provide some assurance in evaluation studies, but the suspicion of imbalance of influential factors cannot be ruled out. Any imbalance can bias the comparison of the treatment groups, if the unbalanced covariates are related to the outcome measure under consideration. The use of propensity score adjustment (by modeling the probability of treatment or sub-classification into groups with similar probability of treatment) eliminates the imbalance of influential factors, which can bias the estimation of treatment effects. In sub-classification, the comparison of treatment groups is conducted within a subgroup, that is defined according to one or more fixed values of covariates related to the probability of treatment self-selection (D'Agostino, 1998).

The method of propensity score adjustment based on sub-classification results in conditional independence of treatment assignment and other covariates. The strategy is to create homogeneous subgroups of subjects with similar predicted probability of selecting a treatment. Models may be developed using any observed covariate information to predict the probability of selecting a treatment. Groups of subjects are identified with similar propensity to select a treatment through sub-classification by influential covariates. The hope is that within these homogeneous groups, there will be little or no imbalance of unobserved factors that might bias treatment comparison. However, the possibility remains that other factors, omitted in the course of data collection, may cause bias in comparing treatments, if predictive of the outcome.

There are several options for within strata analyses and for summary measures of treatment effect. The average treatment effect or odds ratio estimates can be estimated by combining strata. Odds ratio estimates based on logistic regression analysis within strata, to explore or adjust for

additional covariate effects, may be combined into a summary measure. Hypothesis testing is possible using the traditional Mantel-Haenszel approach or an equivalence (or non-inferiority) testing procedure, depending on the study goals and the relative importance of type I and type II errors.

To evaluate the causal effect (Pearl, 2000) of a self-selected treatment on some particular outcome, the causal model involves the effects of measured covariates on the individual subjects' choice of treatment. The following simple map illustrates the causal framework, using propensity score adjustment:

$$Z \rightarrow X \rightarrow Y$$

Where Z is a multidimensional vector of observed covariates, X is the self-selected treatment and Y is the outcome. In this application, Z consists of a number of subject and workplace characteristics determined from baseline interview responses. X is self-reported belt wearing habits (X=0 for subjects who reported "never" wearing back belts; and X=1 for subjects who reported wearing back belts "usually every day"). The outcome, Y, is self-reported low back pain. We want to estimate the average treatment effect, E[Y|X=0] – E[Y|X=1] or the difference in self-reported back pain between the two groups. Through sub-classification, within subgroups, the effect of X → Y is evaluated in a subgroup with Z fixed. The method of classification trees is used to accomplish this goal.

METHODS

Wassell et al. (2000) conducted logistic regression analysis of 6,011 employees to identify significant determinants of self-reported incident back pain. In this causal analysis, the data is restricted, as shown in Table 1, to a group of subjects, who indicated that they had no prior history of back pain or back injury. Having a prior history of back problems was the strongest predictor of back pain in the earlier study – this group was not included here in order to focus the analysis on prevention of new instances of back pain and to eliminate any informal use of the back belt as a therapeutic device. Only subjects who reported "usually everyday" or "never" to a question about belt wearing habits were included in this analysis.

Table 1. Selection Of Subject Interviews For The Propensity Adjusted Causal Analysis Of Back Belts.

9,377 completed baseline interviews.
6,311 completed follow up interviews.
5,427 without any history of previous back injury.
4,569 without "casual" belt wearing (see text).
4,203 subjects without missing covariates.

Table 2. Percent and Regression Results for self-reported back pain incidence in 4,203 Material Handlers.

Variable	No. and percent reporting back pain by group	Adjusted Odds Ratio (95% CI)
Self-Reported Belt Wearing		
Never	273 / 1876 (14.6)	1.0
Usually Every day	333 / 2327 (14.3)	0.9 (0.8 – 1.1)
Store Policy		
Voluntary Belt use	263 / 1973 (13.3)	1.0
Belt Use required	343 / 2230 (15.4)	1.2* (>1.0 – 1.5)
Store Type		
Newly Opened	146 / 1075 (13.6)	1.0
Newly Expanded	460 / 3128 (14.7)	1.0 (0.8 – 1.3)
Frequency of lifting >9kg		
Almost Never	66 / 620 (10.6)	1.0
1-2 times per week	123 / 997 (12.3)	1.2 (0.9 – 1.7)
Usually Every day	417 / 2586 (16.1)	1.8** (1.4 – 2.4)
Job Title		
Dpt. Mgr.	153 / 1200 (12.8)	1.0
Stocker	263 / 1673 (15.7)	1.2 (1.0 – 1.6)
Receiver / Unloader	142 / 1022 (13.9)	1.1 (0.9 – 1.5)
Other	48 / 308 (15.6)	1.3 (0.9 – 1.7)
Job Satisfaction		
Good	266 / 2367 (11.2)	1.0
Poor	340 / 1836 (18.5)	1.7** (1.5 – 2.1)
Smoking Status		
Never	299 / 2250 (13.3)	1.0
Former	86 / 562 (15.3)	1.2 (0.9 – 1.6)
Current	221 / 1391 (15.9)	1.2* (>1.0 – 1.5)
Race		
White	499 / 3431 (14.5)	1.0
Other	107 / 772 (13.9)	1.0 (0.8 – 1.3)
Sex		
Men	250 / 1988 (12.6)	1.0
Women	356 / 2215 (16.1)	1.6** (1.3 – 1.9)
Age		
<25	131 / 853 (15.4)	1.0
25 – 39	271 / 1839 (14.7)	0.9 (0.7 – 1.1)
40 – 54	160 / 1198 (13.4)	0.8* (0.6 --<1.0)
≥55	44 / 313 (14.1)	0.8 (0.6 – 1.2)

Subjects who reported intermediate belt use "once or twice a week" or "once or twice a month" were excluded to eliminate any uncertainty associated with casual belt use. The results of a logistic regression analysis of back pain risk factors for the 4,203 subjects are shown in Table 2. See Wassell et al. (2000) for more details about the variables used in this regression.

Table 2 shows that self reported back pain was more frequent among employees in stores that required belt use, those that reported "usually every day" lifting of items >9 kg, those who reported poor job satisfaction (see Johnston et al., 2002), current smokers and women. In these groups, the odds ratio estimates were significantly greater than unity as shown by the 95% confidence intervals. Employees aged 40 – 54 years were less likely to report back pain as compared to workers < 25 years old. Self reported belt wearing, which is the main variable of interest in this study, was not statistically significant. The incidence of back pain is 14% in this restricted data set, as compared to 17% in the earlier study. Except for the store policy variable, the covariate effects are similar to the findings in the previous analysis.

In the present paper, the problem is approached from a causal analysis perspective, adopting methods to adjust for the non-random self-selection of belt wearing. Using interview responses from a comprehensive questionnaire on health, work, leisure activities, psychosocial factors and belt wearing habits, subjects with similar propensity to wear belts were grouped into homogeneous strata. Following the approach of Stone et al. (1995), and Cook and Goldman (1988) classification tree methods were used to form the homogeneous comparison groups.

The result of the classification tree analysis obtained using S-plus software (S-plus, version 6.0, Insightful Corp), is shown in Figure 1. Store Policy regarding belt use was the most influential factor in determining belt use. The "Belt Use Required Policy" is the traditional policy of these stores, and employees are instructed in proper belt use and lifting techniques when initially hired. Job title is the second most influential factor for determining belt wearing, with employees in the most strenuous jobs with the most frequent lifting (the receiver/unloaders and stockers) tending to report belt wearing "usually every day." Other factors related to reporting "usually every day" belt wearing include those who report exercise habits outside of work, frequent on-the-job lifting, good job satisfaction and White race. Figure 1 shows ten terminal nodes, the percentage reporting "usually every day" belt wearing in each of the nodes and the number (n_i) of employees in each of the nodes. Figure 1 also shows the percentage reporting back pain (BP) in each terminal node, along with the belt effect (BE) or difference in the percent reporting back pain in the "never" belt wearing group minus the "usually every day" belt wearing group.

The percent of employees reporting "usually every day" belt wearing ranges from 21% to 84% across the ten terminal nodes. The least likely to report "usually every day" belt wearing were employees working in stores with a voluntary belt use policy, with the job of Department Manager or Other and who report exercising "almost never" or only "once or twice a week." Employees most likely to report belt wearing "usually every day" worked in stores with a policy requiring belt use, with the job of receiver/unloader or stocker, and reported they were White race.

Table 3 presents the characteristics of the ten strata or belt wearing propensity groups and shows the percent of subjects who reported back pain at the follow up interview for subjects who reported "usually every day" and "never" belt wearing habits. The causal effects or differences of the back pain incidence percentages, by stratum, are also reported in Table 3. The odds ratios and 95% confidence intervals are presented for each stratum. A test of homogeneity of odds ratios across the ten strata (Zelen statistic) was calculated using StatXact5. The exact p-value for the test of homogeneity is 0.77, indicating there is insufficient evidence that odds ratios differ across strata. Given that homogeneity of odds ratios is not an unreasonable assumption, the Mantel-Haenszel estimate of the common odds ratio is 1.09 (95% CI, 0.90 - 1.31), indicating that belt wearing "usually every day" is not associated with any reduction in the odds of self-reported back pain.

Table 3 also includes adjusted odds ratio estimates with 95% confidence intervals, which are nearly identical to the unadjusted results. The adjusted odds ratios are obtained from separate logistic regressions within each stratum. These models include additional covariates on store type (super-stores vs. regular stores), frequency of lifting > 20 lbs at work, job satisfaction, smoking, race, sex and age with the exception that covariates that comprise the strata can not be included in the regression models (e.g. job satisfaction was not included in the regressions for stratum number 4 because all subjects in that stratum reported good job satisfaction).

Within the propensity strata, the difference of two binomial proportions was used as a causal effect estimate to determine if wearing back belts reduces incident back pain (the belt effect or BE shown in Figure 1). A weighted estimate (to account

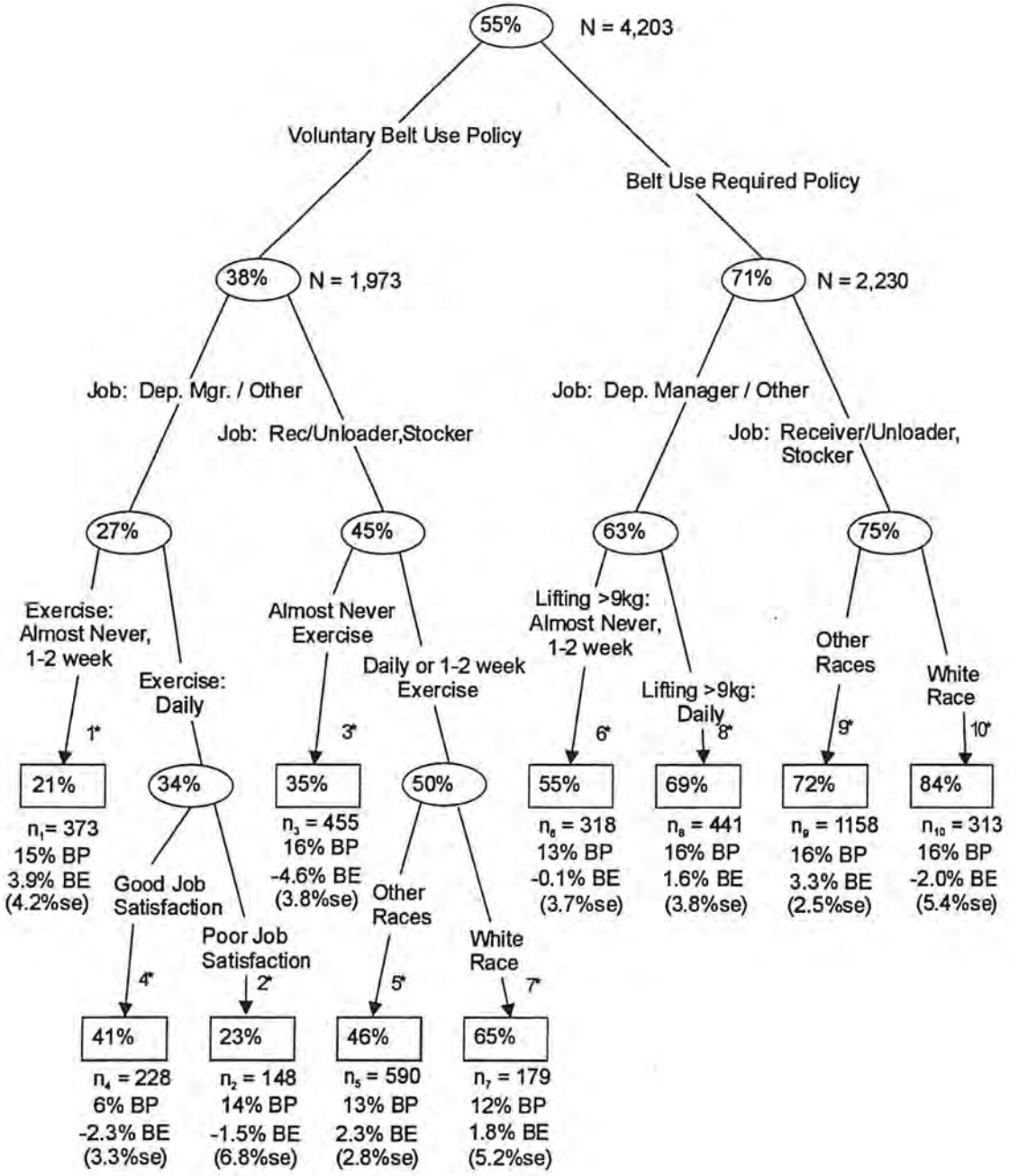

Figure 1. Classification tree for belt wearing showing percent in each node reporting "usually every day" belt wearing habits. Each terminal node is identified by the "starred" numeral for comparison with Table 3. The number of subjects is shown along with the percent reporting back pain at follow up (% BP), belt effect (% BE = p_0-p_1, see Table 1 and text) and the belt effect standard error (%se).

Table 3. Self Reported Back Pain Incidence By Self Reported Belt Wearing And Stratum.

Stratum (Node)	"usually every day" belt wearing			"never" belt wearing					
	No. /w back pain	Total No.	p_1 %	No. /w back pain	Total No.	p_0 %	Odds ratio (95% C.I.)	Adjusted Odds ratio (95% C.I.)	p-value of 1-sided test H_0:BE≥10%
1	9	77	11.7	46	296	15.5	0.7 (0.3-1.6)	0.7 (0.3-1.6)	0.051
2	5	34	14.7	15	114	13.2	1.1 (0.3-3.7)	0.9 (0.3-3.1)	0.023
3	31	157	19.7	45	298	15.1	1.4 (0.8-2.4)	1.5 (0.9-2.5)	<0.001
4	7	93	7.5	7	135	5.2	1.5 (0.4-5.2)	1.6 (0.5-4.9)	<0.001
5	32	271	11.8	45	319	14.1	0.8 (0.5-1.4)	0.8 (0.5-1.3)	0.003
6	22	174	12.6	18	144	12.5	1.0 (0.5-2.1)	0.9 (0.5-1.9)	0.005
7	13	117	11.1	8	62	12.9	0.8 (0.3-2.5)	1.0 (0.4-2.9)	0.074
8	48	303	15.8	24	138	17.4	0.9 (0.5-1.6)	0.9 (0.5-1.6)	0.021
9	124	838	14.8	58	320	18.1	0.8 (0.5-1.1)	0.8 (0.6-1.1)	0.008
10	42	263	16.0	7	50	14.0	1.2 (0.5-3.3)	1.2 (0.5-3.0)	0.032

for differences in stratum sizes) was used to combine the stratum-specific estimates of the belt effects into a summary estimate and confidence interval. The weighted average of the causal effect estimate is 1.0% (95% CI, -6.0% - 8.0%) indicating no evidence that back belts prevent back pain.

In this study, it is important to closely guard against making an incorrect determination that no belt effect exists, if belts are truly effective. A test of non-inferiority, where the type I error rate was set to control the "more serious" error of failing to find a belt effect was used. Specifically, the test was formulated to have a 5% chance of a type I error of declaring belts to be ineffective if there is a ≥10% reduction in back pain incidence. The 10% level of reduction was established in the original study protocol as a level of practical importance. Within each stratum, one-sided tests of H_0: BE ≥ 10% vs. H_A: BE < 10% (exact unconditional tests of non-inferiority using the difference of two binomial proportions) were done using StatXact5 software. The exact p-values were ≤ 0.05 in all but strata 1 and 7; thus 8 out of 10 tests indicated that the belt effect was significantly less than 10%.

Assuming stratum specific tests for the non-inferiority hypothesis are independent, these were combined into a single statistic (Winer, 1971). This calculation involves stratum-specific estimates of $z_i = (BE_i - \Delta)/$ (s.e. rmle)$_i$ using the restricted maximum likelihood standard error (s.e. rmle) estimates appropriate for non-inferiority tests (Miettinen and Nurminen, 1985 implemented in StatXact version 5). The 10 stratum-specific estimates are combined into the sum $\mathbf{Z} = \Sigma\, z_i$ / sqrt(10) , which has a N(0,1) distribution. An iterative process is needed to find the largest value of Δ consistent with the null hypothesis (for which $\Phi(\mathbf{Z}) > 0.05$), because the s.e. rmle depends on the value of Δ. The largest value of a belt effect consistent with a non-inferiority null hypothesis was found to be 2.4%. This is a 95% upper bound for a belt effect from combined stratum-specific, one-sided tests of non-inferiority.

Sensitivity Analysis

Additional support for the full logistic regression (Table 2) is available by conducting a sensitivity analysis to unmeasured confounders (Lin et al., 1998). This interesting alternative approach involves estimating the required strength and imbalance of unobserved effects needed to discredit the observed test results. This approach assumes that greater credibility can be given to results that could only be overturned by a large, prevalent and unbalanced unobserved covariate. If the study is thorough, in collecting a wide range of information and additional covariate information, then it would be unlikely to "miss" a strong, highly influential factor that would have to be very prevalent in the study groups. A study that finds statistically significant but "weak" risk factors demonstrates "thoroughness" in some respects and further substantiates the study results.

To invalidate the results shown in Table 2 (an observed relative risk of 1.0 or no effect of back belts) would require that a risk factor for back pain, with a relative risk of at least 2.0, would have to be more prevalent among the "never" belt wearing group, (with a prevalence of at least 10%). If this risk factor were to exist, it would negate a protective effect of wearing back belts (or a true relative risk of 1.1 for the "never" belt wearing group). In the original study, we searched for risk factors using 151 questionnaire items in a telephone interviews. Table 2 shows that current smokers have a relative risk of 1.2 (statistically significant adjusted odds ratio of 1.2) compared to never smokers. It seems unlikely that a study which measured so many variables (and found significant odds ratios as small as 1.2) would have missed an effect that doubles the risk of back pain

and has a prevalence of ≥10% among the group that reported "never" wearing back belts.

Summary

The methods described in this paper investigate the potential effects of self-selection of treatment. Imbalances of potential confounding factors between the intervention and control groups can be effectively controlled by propensity sub classification methods. Less effective approaches, such as group randomization, cannot eliminate the possibility of invalid results due to confounding. Limitations of this study include the self-reported nature of the data and the potential for selection bias because of the high "turnover" in this type of work. Several analytical strategies described in Wassell et al., 2000 addressed the possible effects of selection bias for the results and other limitations.

In this paper, a causal analysis approach is used to evaluate the effectiveness of back belts to prevent back pain in material handlers. The data excludes subjects with a history of prior back injury and subjects with inconsistent use of the back belt. Self reported back belt use and other risk factors (determined at baseline interview) were evaluated as predictors of self reported back pain (determined at 6 month follow up interview). To adjust for self-selection of treatment, homogeneous subgroups of the data were formed using classification tree methods to predict a subject's choice to wear or not wear back belts. Within these strata, a combined, weighted causal effect estimate (and confidence interval) was estimated that indicated that back belts do not prevent back pain. In order to directly control the possibility of failing to find an effect of back belts, non-inferiority tests were applied in each stratum. A summary test to combine the stratum specific statistics was evaluated and was used to estimate an upper bound estimate of a belt effect. The sensitivity of the results to a potential unobserved and unmeasured risk factor was investigated, which showed the results are unlikely to be invalidated by some unobserved factor.

References

Cook EF and L Goldman. Asymmetric Stratification: An Outline for an Efficient Method for Controlling Confounding in Cohort Studies. Am J Epi. 1988, 127(3): 626-639.

D'Agostino, RB. Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a non-randomized control group. Statist Med. 1998, 17, 2265-2281.

Johnston JM, Landsittel DP, Nelson NA, Gardner LI, Wassell JT. Stressful Psychosocial Work Environment Increases Risk of Back Pain Among Retail Material Handlers. *In Press:* Am J Ind Med. 2002.

Kraus JF, Schaffer KB, Rice T, Maroosis J, Harper J. A Field Trial of Back Belts to Reduce the Incidence of Acute Low Back Injuries in New York City Home Attendants. Int J Occup Environ Health. 2002, 8:9-104.

Lin DY, Psaty BM, Kronmal RA. Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies. Biometrics, 1998, 54: 948-963.

Miettinen OS and M. Nurminen. Comparative Analysis of Two Rates. Statist Med. 1985, 4: 213-226.

Pearl J. Causality: Models, Reasoning and Inference. 2000. Cambridge University Press.

Rosenbaum PR and DB Rubin. The Central Role Of The Propensity Score In Observational Studies For Causal Effects. Biometrika, 1983, 70(1): 41-55.

Rosenbaum PR and DB Rubin. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. JASA. 1984, 79(387): 516-524.

S-Plus (version 6.0). Insightful Corp. Seattle, WA.

StatXact. *A Statistical Package for Exact Nonparametric Inference (version 5.0)*. Cytel Software: Cambridge, MA, 2001.

Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine JM and the Pneumonia Patient Outcome Research Team Investigators. Propensity Score Adjustment for Pretreatment Differences Between Hospitalized and Ambulatory Patients with Community-Acquired Pneumonia. Medical Care. 1995, 33(4): AS56-AS66, Supplement.

Wassell JT. Probability Models for Occupational Injury. Human and Ecological Risk Assessment. 1998, 4(6): 1275-1283.

Wassell JT, Gardner LI, Landsittel DP, Johnston JJ, Johnston JM. A Prospective Study of Back Belts for Prevention of Back Pain and Injury. J. Am. Med. Assoc. 2000, 284(21): 2727-2732.

Winer BJ. Statistical Principles in Experimental Design. 1971. McGraw-Hill Book Co.