Detailed Quantitative Bias Analysis Methods: Injury at Work and Mortality Accounting for Unobserved Smoking and Obesity

We would like to adjust for a missing six-level smoking-obesity variable, but it was not measured in our primary dataset. In order to simulate this six-level variable in our dataset, we need to know the probability of having each level of this six-level variable within levels of the exposure and outcome. In other words, we need to know four sets (lost-time/alive, lost-time/died, medical-only/alive, medical-only/died) of 6 probabilities each of which sum to 100%; that is, the probability of being in each of the six levels of the smoking and obesity variable for each of the four combinations of injury and mortality (indexed 1-6) shown in Table A1. If we knew this, then we could simulate each person in the dataset's confounder status using the appropriate probabilities. For example, for a person who had a lost time injury and died, we could choose a draw from a random uniform distribution between 0 and 1 and if the value was between 0-p1, we would assign them to be an obese current smoker, between p1 and p2, an obese former smoker, between (p1+p2) and p3, an obese never smoker, between (p1+p2+p3) and p4, a non-obese current smoker, between (p1+p2+p3+p4) and p5, a non-obese former smoker and above (p1+p2+p3+p4+p5) a non-obese never smoker. We could do the same for each of the four combinations of lost time and mortality groups. However, we do not know these values, and we cannot simply put distributions on them without more information on how the two confounders (smoking and obesity) relate to both the exposure and the outcome.

To simulate the data that would have occurred had we collected data on the confounders, we combine the observed data, the distribution of lost time injury and mortality shown in the top of Table A2, and estimates of the relationship between the unmeasured confounder and both the lost-time injury and the mortality generated from external data sources. Here we know the A, B, C and D cells but do not know how these values are distributed with respect to smoking and obesity (i.e., the subscripted cells in Table A3). If we could know the values in Table A3, we could complete the values in Table A1. For example, p1, the probability of being an obese current smoker given one had a lost time injury and died is now just A_1/A or p(Obese, Current Smoker|Died, Lost Time Injury)=A_1/A with A being known and A_1 being unknown. So, to generate data on the probabilities in Table A1, we need to complete Table A3. The bottom panel of Table A2 shows the actual data presented in Table 1 of the paper, only among women.

To complete Table A3, we can use information on the distribution of obesity/smoking within injury type (which would allow us to complete the N_i and M_i cells of the table) and the strength of the effect of each level of smoking and obesity on mortality (which would allow us to complete the A_i and B_i cells of the table, leaving the C_i and D_i cells known given the rest of the values in the table). If we know these or, as in this case, we can make reasonable assumptions about what those parameters (known as bias parameters) are, we can complete Table A3, which would allow us to complete Table A1.

To make reasonable inferences about the distribution of the obesity/smoking variable within injury types, we used data from the Panel Study of Income Dynamics (PSID), which has information on smoking and obesity and lost time injuries. While this is not the same as our primary dataset, we believe it provides a reasonable approximation to the likely distribution (with error as we will account for later) of obesity and smoking in our population. Thus, we used the observed distributions in the PSID data to estimate the distribution of obesity/smoking within injury types in our dataset. Knowing this, allows us to obtain the subscripted M and N variables in Table A3 by multiplying the probabilities we estimated from the PSID by the total M and N numbers in our primary data. Table A4 presents the distribution of obesity and smoking within levels of injury type for women in the PSID data that we used to complete part of Table A3. For women, this would allow us to complete part of Table A3 as in Table A5. As an example, we estimated that among women, 48.2% of the uninjured were non-obese, never smokers. Given in our main dataset we had 25,980 women without lost-time injuries; that equates to 12,522.36 non-obese, never smokers in this group.

Now that we have the denominators completed, we move on to the interior cells of the Table A3. To determine the subscripted A-D cells in this table we need to know the strength of the effect (defined as a relative risk) of each level of the confounder on mortality. Again, these are not available in our dataset but we can estimate these from regressions using the National Health and Nutrition Examination Survey (NHANES) data, described in more detail in the main text. As with the PSID data, this is not a perfect comparison, but we think a reasonable approximation. Table A6 shows the relative effect sizes. For example, we estimate the mortality effect of being a current smoker and obese compared to never smoker and not obese is 4.26 (95% CI 3.41-5.33). We will want to recreate this relationship within our dataset.

We can do this because knowing the strength of effect of obesity/smoking on mortality gives us a series of relationships between the confounder (vs. a baseline level of a non-obese, never smoker) and the outcome. For example, the mortality risk ratio for obese/current smokers vs non-obese/never smokers among the lost-time injured workers (which we estimate is 4.26) is equal to $(A_1/M_1)/(A_6/M_6)$ in our dataset and equal to $(B_1/N_1)/(B_6/N_6)$ among the comparison workers without lost-time injuries. Note that we now know each of the subscripted M and N cells from the previous step. To determine the values of each of the A and B cells, one could solve for the unknowns given the PSID smoking and obesity prevalence by sex and injury type and the NHANES mortality hazard ratios by smoking and obesity. However, as there are so many parameters, this is a problem that we can solve more easily through simulation. We simulated values for each of the A and B cells and selected the combination of cell counts that minimized the difference between the measured relative risk values in the PSID simulation and the expected relative risk values based on the NHANES regression (Table A6) and on the unconditional relationship between injury and mortality reported in the earlier study. This allowed us to identify values for all cells in Table A3 that were reasonable approximations of the values in the NHANES and PSID data in terms of the distribution of the confounder with respect to the exposure (lost time injury) and the confounder's strength of effect on our outcome (mortality). We show an example of a simulated dataset that meets the specifications in Table A7. We rearrange this dataset in Table A8 to be grouped by the obesity and smoking groups, which show that the resulting simulated relationships are very similar to what we desired in Table A6.

The methods above assume that we know the effect of smoking and obesity in our dataset with certainty, which we do not. Since we estimated the relative risk values for smoking and obesity with uncertainty in the NHANES regressions they may not perfectly transport to our dataset, we repeated the entire process described above over 100 iterations, each iteration sampling random hazard ratios for each smoking and obesity group based on normal distributions based on the regression results and standard errors. Each time this gave us new values for Table A2, which then changed the values in Table A3.

Once we had Table A3 completed, we could now calculate the probability of having each level of the confounder within levels of the exposure and outcome as shown in Table A1. For example, the probability of being obese and a current smoker among those who had lost-time injuries and died is A_6/A, and among those without exposure and outcomes is D_6/D. Again, note that the probabilities for each of the six subscripted versions of each letter must add up to 100%.

After doing this for each of the A-D cells, we simulated the confounder within our primary dataset using random draws from a uniform distribution and assigning a person to one of the six levels of the confounder depending on their actual exposure and outcome status with a probability equal to the estimated probability from the previous analysis. For example, assume that there is a 10% probability that an observation in the lost-time injury and mortality-event group is a non-obese never smoker. We assign this obesity-smoking category to any observations that get a uniform random draw between 0.0 and 0.1. We assign the other five obesity-smoking groups similar intervals in the uniform distribution probability space.

The approach just described gives us a single simulated confounder for each person in the primary dataset that we could use to adjust for smoking and obesity. However, because there was uncertainty in the estimates of those probabilities, rather than using the estimated probabilities as fixed parameters, we assigned distributions to each probability and randomly sampled from those distributions. Specifically, we used normal distributions around the logit transformations of the calculated probabilities. We calculated the standard deviation of the normal distributions as the logit transformation of a proportion estimate from samples corresponding to the size of the relevant group of observations in the PSID data. To ensure valid probability distributions within each lost-time injury and mortality-event group, we scaled the drawn probabilities by dividing by the sum of the drawn probabilities. Doing this once gave us a single possible result given the uncertainty in the distributions and allowed us to calculate a hazard ratio for the effect of lost time injury on mortality adjusted for smoking and obesity (in addition to measured confounders). As this is only one possible result we could have gotten, we then repeated this process 10,000 times (for 100 simulations of smoking and obesity values in the primary data for each of the 100 probability distributions obtained from the PSID calibration) and simulated 10,000 adjusted hazard ratios.

The methodology above aims to adjust for systematic error from the smoking and obesity confounders. For each of the 10,000 simulated hazard ratios, we also simulated random error using an error draw from a mean zero normal distribution with a standard deviation based on the standard error of the unadjusted hazard ratio estimate. We summarize the 10,000 adjusted hazard ratios using the median as the point estimate and the 2.5th to 97.5th percentiles as a simulation interval describing the totality of results adjusted for the missing confounders.

Table A1 – Probability of having each level of the 6-level unmeasured confounder (smoking and obesity) within levels of lost-time injuries and death.

Group	Probability
Lost-time injury, died	
Obese, current smoker	p1
Obese, former smoker	p2
Obese, never smoker	p3
Non-obese, current smoker	p4
Non-obese, former smoker	p5
Non-obese, never smoker	p6
No lost-time injury, died	
Obese, current smoker	q1
Obese, former smoker	q2
Obese, never smoker	q3
Non-obese, current smoker	q4
Non-obese, former smoker	q5
Non-obese, never smoker	q6
Lost-time injury, alive	
Obese, current smoker	r1
Obese, former smoker	r2
Obese, never smoker	r3
Non-obese, current smoker	r4
Non-obese, former smoker	r5
Non-obese, never smoker	r6
No lost-time injury, alive	
Obese, current smoker	s1
Obese, former smoker	s2
Obese, never smoker	s3
Non-obese, current smoker	s4
Non-obese, former smoker	s5
Non-obese, never smoker	s6

Table A2 – Crude hypothetical (top) and actual (bottom) data on the relationship between lost time injuries and mortality for women, primary dataset

	Died	**Survived**	**Total**
Lost-time injury	A	C	N
Medical-only injury	B	D	M
	Died	**Survived**	**Total**
Lost-time injury	936	11,421	12,357
Medical-only injury	1,265	24,715	25,980

Table A3 – Crude data on the relationship between lost time injuries and mortality stratified by obesity and smoking for women

	Died	**Survived**	**Total**
Total			
Lost-time injury	936	11,421	12,357
No lost-time injury	1,265	24,715	25,980
Obese, current smoker			
Lost-time injury	A_1	C_1	N_1
No lost-time injury	B_1	D_1	M_1
Obese, former smoker			
Lost-time injury	A_2	C_2	N_2
No lost-time injury	B_2	D_2	M_2
Obese, never smoker			
Lost-time injury	A_3	C_3	N_3
No lost-time injury	B_3	D_3	M_3
Not-obese, current smoker			
Lost-time injury	A_4	C_4	N_4
No lost-time injury	B_4	D_4	M_4
Not-obese, former smoker			
Lost-time injury	A_5	C_5	N_5
No lost-time injury	B_5	D_5	M_5
Not obese, never smoker			
Lost-time injury	A_6	C_6	N_6
No lost-time injury	B_6	D_6	M_6

Table A4. PSID Distribution of Obesity and Smoking within Levels of Injury Type for Women

	Women		
	N	Lost-Time Injury (Weighted %)[a]	Uninjured (Weighted %)[a]
Sample	2,457	100.0%	100.0%
Obese, current smoker	163	6.1%	5.4%
Obese, former smoker	82	4.3%	3.8%
Obese, never smoker	504	24.8%	15.4%
Non-obese, current smoker	442	26.6%	19.1%
Non-obese, former smoker	165	6.2%	8.2%
Non-obese, never smoker	1,101	32.0%	48.2%

[a] Percentages in table use PSID sampling weights.
Note: This Table is drawn from Table 2 in the paper.

Table A5 – Crude data on the relationship among women between lost time injuries and mortality stratified by obesity and smoking using the PSID data to complete the totals

	Died	Survived	Total	
Total				
Lost-time injury	936	11,421	12,357	
No lost-time injury	1,265	24,715	25,980	
Obese, current smoker				
Lost-time injury	A_1	C_1	12,357*6.1%	=757.48
No lost-time injury	B_1	D_1	25,980*5.4%	=1,395.13
Obese, former smoker				
Lost-time injury	A_2	C_2	12,357*4.3%	=525.17
No lost-time injury	B_2	D_2	25,980*3.8%	=974.25
Obese, never smoker				
Lost-time injury	A_3	C_3	12,357*24.8%	=3,065.77
No lost-time injury	B_3	D_3	25,980*15.4%	=4,006.12
Not-obese, current smoker				
Lost-time injury	A_4	C_4	12,357*26.6%	=3,289.43
No lost-time injury	B_4	D_4	25,980*19.1%	=5,012.00
Not-obese, former smoker				
Lost-time injury	A_5	C_5	12,357*6.2%	=759.96
No lost-time injury	B_5	D_5	25,980*8.2%	=2,125.16
Not obese, never smoker				
Lost-time injury	A_6	C_6	12,357*32.0%	=3,957.95
No lost-time injury	B_6	D_6	25,980*48.2%	=12,524.96

Note: We calculated the counts within obesity/smoking groups using the unrounded percentages from Table A4; there may be discrepancies between these values and the products that use the rounded percentages presented above.

Table A6. Regression Results: Mortality of Non-Federal Wage and Salary Workers (N=14,509)

Baseline characteristics	Hazard Ratio	95% Confidence Interval	
Sex			
Women	Ref	--	--
Men	1.32	1.18	1.49
Race/ethnicity			
White	Ref	--	--
Black	1.38	1.20	1.57
Hispanic	1.36	1.09	1.79
Other	1.25	0.82	1.91
Place of birth			
Not U.S. Born	Ref	--	--
U.S. Born	0.90	0.75	1.10
Education			
Less than high school	Ref	--	--
High school or equiv.	1.11	0.94	1.30
More than high school	1.08	0.89	1.30
Smoking-Obesity Category			
Obese, current smoker	4.26	3.41	5.33
Obese, former smoker	1.99	1.55	2.57
Obese, never smoker	1.50	1.2	1.89
Not obese, current smoker	2.90	2.39	3.52
Not obese, former smoker	1.21	0.96	1.51
Not obese, never smoker	Ref	--	--

Table A4 presents mortality hazard ratios estimated from NHANES III (1988-1994) and continuous waves (1999-2014), including people ages 35-74 at the time of the survey who were employed in the private sector or state and local government.

Table A7 – Crude simulated data on the relationship among women between lost time injuries and mortality stratified by obesity and smoking using the PSID data to complete the interior cells

	Died	Survived	Total
Total			
Lost-time injury	936	11,421	12,357
No lost-time injury	1,265	24,715	25,980
Obese, current smoker			
Lost-time injury	129.46	627.97	757.43
No lost-time injury	173.34	1,221.79	1,395.13
Obese, former smoker			
Lost-time injury	41.99	483.03	525.02
No lost-time injury	56.60	917.65	974.25
Obese, never smoker			
Lost-time injury	185.20	2,880.57	3,065.77
No lost-time injury	175.62	3,830.50	4,006.12
Not-obese, current smoker			
Lost-time injury	383.55	2,906.50	3,290.05
No lost-time injury	419.23	4,535.16	4,954.39
Not-obese, former smoker			
Lost-time injury	36.92	723.52	760.44
No lost-time injury	75.22	2,049.94	2,125.16
Not obese, never smoker			
Lost-time injury	158.91	3,799.41	3,958.32
No lost-time injury	365.00	12,159.96	12,524.96

Table A8 – Table A7 rearranged to be grouped by obesity and smoking to demonstrate the simulated relationship between smoking and obesity and mortality in the primary dataset compared to the desired relationships as presented in Table A6

	Died	Survived	Total	Risk ratio	Simulated risk ratio
Total					
Lost-time injury	936	11,421	12,357		
No lost-time injury	1,265	24,715	25,980		
Lost-time injury					
Obese, current smoker	129.46	627.97	757.43	4.26	4.93
Obese, former smoker	41.99	483.03	525.02	1.99	2.08
Obese, never smoker	185.20	2,880.57	3,065.77	1.50	1.54
Not-obese, current smoker	383.55	2,906.50	3,290.05	2.90	3.16
Not-obese, former smoker	36.92	723.52	760.44	1.21	1.22
Not obese, never smoker	158.91	3,799.41	3,958.32	1.00	1.00
No lost-time injury					
Obese, current smoker	173.34	1,221.79	1,395.13	4.26	4.73
Obese, former smoker	56.60	917.65	974.25	1.99	2.05
Obese, never smoker	175.62	3,830.50	4,006.12	1.50	1.53
Not-obese, current smoker	419.23	4,535.16	4,954.39	2.90	3.08
Not-obese, former smoker	75.22	2,049.94	2,125.16	1.21	1.22
Not obese, never smoker	365.00	12,159.96	12,524.96	1.00	1.00