Bayesian modelling of lung cancer risk and bitumen fume exposure adjusted for unmeasured confounding by smoking

F de Vocht, H Kromhout, G Ferro, P Boffetta and I Burstyn

Occup. Environ. Med. 2009;66;502-508; originally published online 5 Dec 2008; doi:10.1136/oem.2008.042606

Bayesian modelling of lung cancer risk and bitumen fume exposure adjusted for unmeasured confounding by smoking

F de Vocht,[1,2] H Kromhout,[3] G Ferro,[2] P Boffetta,[2] I Burstyn[4]

[1] Occupational and Environmental Health Research Group, School of Translational Medicine, Faculty of Medical and Human Sciences, The University of Manchester, Manchester, UK; [2] International Agency for Research on Cancer (IARC), Lyon, France; [3] Environmental Epidemiology Group, Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands; [4] Community and Occupational Medicine Program, Department of Medicine, Faculty of Medicine and Dentistry, The University of Alberta, Edmonton, Canada

Correspondence to: Frank de Vocht, Occupational and Environmental Health Research Group, School of Translational Medicine, Faculty of Medical and Human Sciences, The University of Manchester, Manchester M13 8GE, UK; frank.devocht@manchester.ac.uk

Accepted 7 November 2008
Published Online First 5 December 2008

ABSTRACT

Objectives: Residual confounding can be present in epidemiological studies because information on confounding factors was not collected. A Bayesian framework, which has the advantage over frequentist methods that the uncertainty in the association between the confounding factor and exposure and disease can be reflected in the credible intervals of the risk parameter, is proposed to assess the magnitude and direction of this bias.

Methods: To illustrate this method, bias from smoking as an unmeasured confounder in a cohort study of lung cancer risk in the European asphalt industry was assessed. A Poisson disease model was specified to assess lung cancer risk associated with career average, cumulative and lagged bitumen fume exposure. Prior distributions for the exposure strata, as well as for other covariates, were specified as uninformative normal distributions. The priors on smoking habits were specified as Dirichlet distributions based on smoking prevalence estimates available for a sub-cohort and assumptions about precision of these estimates.

Results: Median bias in this example was estimated at 13%, and suggested an attenuating effect on the original exposure–disease associations. Nonetheless, the results still implied an increased lung cancer risk, especially for average exposure.

Conclusions: This Bayesian framework provides a method to assess the bias from an unmeasured confounding factor taking into account the uncertainty surrounding the estimate and from random sampling error. Specifically for this example, the bias arising from unmeasured smoking history in this asphalt workers' cohort is unlikely to explain the increased lung cancer risk associated with average bitumen fume exposure found in the original study.

Even though in epidemiological studies as much care as possible is normally taken to collect information on factors that might potentially confound the association between the exposure of interest and the disease of interest,[1–3] for various reasons sometimes this information has not been or could not have been collected.[4] Specifically, occupational cohort studies typically lack data on risks that arise from an individual's behaviour which are not recorded in sources of information normally used to construct the cohorts, such as records kept by the employer or exposure measurements. As a result, residual confounding may bias exposure–disease associations reported in such studies. Additionally, it has been shown that residual confounding can still be present after controlling for a confounder that is imprecisely measured.[5] Because the amount of bias from these sources is unknown and bias is expected to be directional, the frequentist confidence intervals tend to underestimate the uncertainty about the true effect[6] and may even present biased estimates of such uncertainty.

Several methods to estimate the sensitivity of the risk estimates to unmeasured confounders have been proposed,[7–10] but they generally require information on the prevalence of the unmeasured confounder, its association with the exposure and its effect on the outcome[11] to "externally adjust" the original risk estimate. Conventional sensitivity analyses are also difficult to summarise as the number of parameters determining the bias increases and does not provide a full range for likely bias in the results, and can further be misleading since the probability of the presented scenarios is usually not taken into account.[6]

A Bayesian framework can be applied to improve sensitivity analyses by incorporating uncertainty regarding bias in the results,[6 12–14] which has the advantage over ordinary sensitivity analyses that the uncertainty regarding the relationship of the confounding variable to exposure and disease can be reflected in credible intervals of the distribution of true values of the risk parameter,[6 12] especially when some information is available on the unmeasured confounding factor. In this paper we adapt a Bayesian adjustment method proposed by Steenland and Greenland,[6] extending it to a full Bayesian framework to produce a distribution of risk parameters reflecting the uncertainty due to both random sampling error and the uncertainty about the unmeasured confounder.

To illustrate this extended method we will use it to assess bias from smoking as an unmeasured confounder in a cohort study of lung cancer risk from bitumen fume exposure in the European asphalt industry. The association between occupational exposure to bitumen fume during road paving and roofing and excess lung cancer risk has been under discussion for a long time.[15–22] One reason for this is the uncertainty about the impact of residual confounding on the results of the published studies from co-exposures to other potential carcinogens[19] and lifestyle factors, including smoking of tobacco products.[17] We will assess data from a multi-centre cohort study that provided evidence that excess lung cancer risk is associated with employment in the asphalt industry[20] and may in particular be associated in an exposure-dependent manner with career average exposure to bitumen fume.[21] Unfortunately, it was

not feasible in the original cohort to collect individual smoking histories,[20 21] even though smoking is considered the main cause of lung cancer[22] and thus has a potential for being an important confounding factor in the study.[17 18]

There is an opportunity to "externally" adjust the risk estimates in the asphalt cohort that is the subject of this paper for smoking, since individual smoking history was collected in 31% of the Dutch sub-cohort (n = 1138 workers) who underwent routine medical examinations.[16] Internal sensitivity analyses by Hooiveld and colleagues[16] that use an adjustment method described by Axelson and Steenland,[9] found little potential for confounding in the Dutch sub-cohort. However, these analyses suffered from simplistic assumptions about the potential distribution of smoking habits and did not propagate uncertainty about the (unknown) distribution of smoking habits to estimates of exposure–disease associations. The uncertainty about the extent to which observed smoking habits reflect those of typical cohort members was not considered. To address these limitations of previous work, we have conducted a fully Bayesian sensitivity analysis of the confounding by latent smoking habits of the association between bitumen fume and lung cancer risk in the international cohort. We first describe the cohort in greater detail and present the statistical model we adopted. Finally, we present the result of Bayesian sensitivity analysis for confounding and discuss the results.

METHODS

Cohort of European asphalt workers

The study population has been described in detail.[20 21] In summary, it included 79 822 male workers employed between 1913 and 1999 for at least 1 year in the asphalt industry in Denmark, Finland, France, Germany, Israel, the Netherlands, Sweden and Norway. Follow-up differed between countries and started between 1953 and 1979 and ended between 1995 and 2000.

The Swedish sub-cohort was excluded from the analyses presented here because individual employment histories were absent. Quantitative exposure estimates were derived for cohort members only employed in asphalt paving using a study-specific exposure matrix,[23 24] and as such these analyses included 12 367 male asphalt workers of whom 135 had died due to lung cancer as the primary cause of death.

Subject-specific exposure estimates were separated into four exposure groups in such a way as to ensure equal numbers of lung cancer cases in each stratum.[23] The lowest exposed group was used as the reference group in the statistical models that estimated relative risks.

Disease model

The frequentist method used to assess the association between average bitumen fume exposure and lung cancer risk in the original analyses[21] applied multiple Poisson regression models to adjust for potential confounding by country, calendar year, age, duration of employment, and the use of coal tar. This disease model did not take into account confounding by smoking:

$$Y_i \sim \text{Poisson}(\mu_i), \text{ with} \quad (1)$$

$$\log(\mu_i) = \alpha_0 + \text{offset}_{\ln(\text{personyears})i} + z_J * \text{bitumen fume exposure}_i + \delta_{1-9} * \text{age}_i + \delta_{10-13} * \text{year}_i + \delta_{14-17} * \text{duration of employment}_i + \delta_{18-23} * \text{country}_i + \delta_{24} * \text{use of coaltar}_i$$

where α_0 is the intercept, j the exposure levels 1 to J, where J = 4 for unlagged exposures and J = 5 for lagged exposures. We assume that the true disease model also includes a term for tobacco smoking.

Latent confounder model

Steenland and Greenland[6] proposed two methods to conduct sensitivity analysis using information on the distribution of a confounding factor from external sources: (a) a method based on independent Monte Carlo sampling from scenarios based on a prior distribution of the confounding factor in an external population, and (b) a Bayesian method in which the prior distribution is combined with observed data and sampling is done from the combined posterior distribution. We adopted and extended their Bayesian sensitivity analysis approach to assess the amount of bias associated with a particular confounding factor (smoking in their and our example), which relies on the availability of an a priori estimate of the prevalence of this confounding factor in the population under investigation.

The underlying model assumes that relative risks (RR) comparing the exposure-smoking-specific group with referent non-smokers within covariate strata, assuming that the rate-ratios do not vary across covariate levels, can be described as:

$$RR_{\text{exposure–smoking–specific}} = e^{Z_j x_j + \gamma_1 \mu_1 + \gamma_2 \mu_2} \quad (2)$$

where e^{Z_j} is the RR relating exposure to disease risk within specific levels of (bitumen fume) exposure, while e^{γ_1} and e^{γ_2} are the RRs for the unmeasured confounding factor: current and former smokers versus never smokers, respectively (represented by dummy variables (1/0) μ_1 and μ_2). Subsequently, bias in the jth exposure stratum can be estimated as (here we suppress subscript j for simplicity):

$$\text{Bias} = \frac{RR_{\text{exp }0}}{RR_{\text{non exp }0}} = \frac{P_{\text{never, exp}} + e^{\beta_2} P_{\text{current, exp}} + e^{\beta_3} P_{\text{former, exp}}}{P_{\text{never, non exp}} + e^{\beta_2} P_{\text{current, non exp}} + e^{\beta_3} P_{\text{former, non exp}}} \quad (3)$$

where $p_{\text{never,exp}}$, $p_{\text{current,exp}}$, $p_{\text{former,exp}}$, $p_{\text{never,nonexp}}$, $p_{\text{current,nonexp}}$ and $p_{\text{former,nonexp}}$ are the proportions of never, current and former smokers in the exposed and referents, or in this example in the different average bitumen fume exposure strata. Assuming that bias is approximately constant across strata, the confounder-adjusted RR for jth exposure is calculated by:

$$\theta_j = RR_{\text{adjusted}} = \frac{RR_{\text{unadjusted}}}{\text{bias}} \quad (4)$$

When there are a priori reasons to assume that the proportions of smokers might differ between strata with different exposure levels, this model can easily be extended by incorporating strata-specific exposure-smoking-specific RRs, which will also be illustrated in this example.

Posterior

Given the statistical model and adjustment factor for smoking as an unmeasured confounder, the posterior summarisation can be described, given p[.] as the probability density function of a random variable, Z as the vector of the Poisson regression parameters (z_j) of log-relative risks due to bitumen fume exposure before adjustment for the unmeasured confounder, θ as the vector of log-relative risks adjusted for unmeasured

confounder, G is the matrix defining membership in a bitumen fume exposed group, Y is the Poisson-distributed count of lung cancers in each strata, D is the vector of the Poisson regression parameters (δ) for measured confounders (including intercept α_0), Γ is the vector of constants in γ_1 and γ_2 used in the calculation of bias due to unmeasured confounder, and π is the vector of prevalence of unmeasured confounder (smoking) for different exposure groups, as:

$$p[Z, \delta, \theta, \pi \mid Y, G, D, \Gamma] \propto$$

$$p[Y \mid Z, D, G]\, p[\theta \mid Z, \pi, \Gamma]\, p[Z]\, p[D]\, p[\theta]\, p[\pi] \quad (5)$$

Priors

Prior distributions for effect of current (γ_1) and former (γ_2) smoking on risk of lung cancer were specified as normal distributions with means and variance based on the odds ratios (OR) for current cigarette smoking (OR 23.9, 95% confidence interval 19.7 to 29.0) and former cigarette smoking (OR 7.5 (6.2 to 9.1)) among European men,[25] which were obtained from a study combining data from 10 case–control studies from six European countries and included 7609 cases of lung cancer and 10 431 controls. The priors for the intercept (α_0), the estimated effects of exposure (z_j) and the measured confounders (δ_{1-4}) were specified as non-informative Gaussian distributions with mean 0 and variance 10 000.

The proportions of never, former and current smokers in the exposed (21%, 32% and 47%, respectively) and referent populations (28%, 32% and 40%, respectively) in this cohort were obtained from the data for the Dutch sub-cohort,[16] the assumption being that the sub-cohort has a similar control population and similar demographics as the complete cohort. Furthermore, we make a rather anti-conservative assumption that the prevalence of smoking in the low exposed stratum is similar to that in a non-exposed referent population, whereas in fact the difference in smoking habits among them and highly exposed asphalt pavers is thereby exaggerated, possibly leading to over-correction for the unmeasured confounder. We assume informative priors for these proportions based on a Dirichlet distribution. Steenland and Greenland[6] discussed the Dirichlet distribution as the valid distribution to be used in these situations but used a more familiar bivariate normal distribution as an approximation to the logit of the Dirichlet distribution instead. However, with the currently available software the use of this distribution does not pose problems anymore and hence we advocate its use. The Dirichlet distribution can be regarded as the multivariate generalisation of the beta distribution and has been described by Connor and Mosimann.[26] As an example for readers not familiar with this distribution, it has been described on http://en.wikipedia.org/wiki/Dirichlet_distribution as "one example use of the Dirichlet distribution is if one wanted to cut strings (each of initial length 1.0) into K pieces with different lengths, where each piece had, on average, a designated average length, but allowing some variation in the relative sizes of the pieces. The α/α_0 values specify the mean lengths of the cut pieces of string resulting from the distribution. The variance around this mean varies inversely with α_0". Therefore, given that current, former and never-smokers at any point in time form mutually exclusive categories, it is natural to imagine these three proportions follow the Dirichlet distribution.

To illustrate this method, we conducted two series of analyses, assuming 5% ($p_{exposed}$~Dirichlet(α[2.52, 3.84, 5.64]) and $p_{referent}$~Dirichlet (α[3.36, 3.84, 4.80]) and 10% variance ($p_{exposed}$~Dirichlet(α[1.13, 1.70, 2.49]) and $p_{referent}$~Dirichlet (α[1.57, 1.79, 2.24]) in the estimated proportions of never, former and current smokers in the exposed and referent populations. These variance estimates have been chosen as plausible values since the variance of the proportions of smokers in the Dutch cohort were not reported in the original publication.[16] The 5% and 10% have been chosen such that they reflect our a priori uncertainty about the Dutch subsample being a true random sample from the full cohort, while a higher uncertainty does not provide interpretable results due to overly wide credible intervals in the sample from the posterior (not shown). These Dirichlet input values can be calculated using equations 6–9:

$$\text{Dirichlet}(\boldsymbol{\alpha}) \sim (X_{\text{never smoker}}, X_{\text{former smoker}}, X_{\text{current smoker}}) \quad (6)$$

$$X_i \sim \text{beta}(\alpha_i, \alpha_0 - \alpha_i) \quad (7)$$

$$\alpha_0 = \sum_{i=1}^{K} \alpha_i \quad (8)$$

$$E_{x_i} = \frac{\alpha_i}{\alpha_0} \quad (9)$$

$$\text{var}_{x_i} = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \quad (10)$$

Additionally, a second a priori hypothesis was assessed in which we not only assumed an association between smoking prevalence exposure (similar to the previous example) but also assumed that this association increased with increased average exposure to bitumen fume. Although this association has been discussed anecdotally, there is at present no evidence to support this and these simulations as such are for illustrative purposes. This was done by extending the Bayesian framework by calculating three bias factors for the proportions of current and never smokers in the low (42% and 26%), medium (47% and 21%) and high (52% and 16%) exposed strata, while these proportions in the referent stratum remained 40% and 28%, respectively, which was based on an a priori defined 5% decrease in smoking prevalence in the lowest exposed group and a 5% increase in smoking prevalence in the highest exposed group compared to the original estimates.[16]

SAMPLING FROM THE POSTERIOR

Following standard Bayesian methods, numerical results are based on 50 000 samples from the posterior after a 5000 sample "burn in", using WINBUGS 1.4.3.[27] The model syntax is described in Appendix A. The marginal posterior distributions were generated using Gibb's sampling with the Metropolis Markov chain Monte Carlo (MCMC) algorithm, tuned to obtain an acceptance rate between 20% and 40%. Three MCMC chains were obtained simultaneously, using different initial starting values, obtained from the mean and 95% confidence limits from the frequentist Poisson regression model, to test convergence. Convergence was acceptable when Monte Carlo (MC) errors were smaller than 5% and residual correlation in the MCMC chains was absent. We computed 95% credible intervals (2.5th to 97.5th percentile of the posterior distribution) of bias and smoking-adjusted relative risk. Bias, unadjusted relative risks (RRs) and RRs adjusted for confounding by

smoking were estimated at every iteration and summarised over all 50 000 samples using means and 95% credible intervals.

In sampling from the posterior, at each MCMC iteration we drew (a) a sample of risk parameter not adjusted for smoking using non-informative priors on covariates and cohort data in the Poisson disease model, (b) a sample of distributions of smoking habits from the Dirichlet prior on smoking habits, (c) a sample of effects of smoking habits on risk of lung cancer from respective prior distributions, and finally combined the values obtained in steps (a) to (c) in calculating bias-adjusted RR according to equations 2–4. This is different from the method of Steenland and Greenland[6] in which only sampling steps (b) and (c) were repeated for a fixed value of risk parameter. Consequently, in our method credible intervals reflect uncertainly due to both random error in the disease (Poisson) model and the uncertainly in the distribution and effect of the confounder. This constitutes a major methodological advance and makes our method, unlike that of Steenland and Greenland,[6] particularly suitable for internal risk comparisons within cohorts.

RESULTS

The results of the sensitivity analyses are reported in tables 1, 2 and 3. Estimated mean bias differs from median bias depending on the level of uncertainty we a priori assigned to the posterior distribution of smoking prevalence in the population. Nonetheless, although mean bias has been estimated at 19–20% (5% variance) and 27–28% (10% variance), median bias was 13–14% for both simulations, attenuating the exposure–disease associations for average, cumulative and lagged exposure to bitumen fume. However, the 95% credible intervals for bias factors included negative values, implying the possibility of correction away from the null due to some constellations of the distribution of smoking habits in the cohort.

For average exposure the adjusted RRs in the exposed strata ranged from 1.78 (95% credible interval 0.75 to 3.55) in the low exposed group to 2.61 (1.02 to 5.44) in the highest exposed group, respectively, when we are fairly sure about the prevalence estimates (5% variance), and ranged from 1.88 (0.56 to 4.61) to 2.76 (0.79 to 6.95) with larger uncertainty (10% variance). For the other exposure metrics exposure–response associations were found with an RR of 2.22 (5% variance) and 2.35 (10% variance) in the stratum with highest cumulative exposure and RRs of 1.80 (5% variance) and 1.91 (10% variance) and 1.34 (5% variance) and 1.42 (10% variance) in the highest 15-year lagged average and cumulative exposure strata, respectively, with 95% credible intervals of these estimates including unity.

Table 3 shows the results, for average bitumen fume exposure only, when assuming a positive association between level of exposure and smoking prevalence, and shows that median bias ranges from 3.5% (−58.5% to 116.4%) to 3.3% (−58.5% to 116.4%) in the low exposed group to 22.1% (−32.0% to 132.3%) and 22.7% (−45.9% to 145.1%) in the highest exposed stratum. As expected this further attenuates the exposure–effect association to RRs of 2.39 and 2.52 in the highest exposed groups, depending on the level of uncertainty in the smoking prevalence, with the 95% credible intervals now including unity.

DISCUSSION

These analyses suggest that although confounding from smoking, caused by the diversity of smoking habits in the asphalt industry, could have been present in the cohort, the data still support, albeit with reduced certainty, an exposure–response association between average bitumen fume exposure and increased lung cancer risk.

Although estimated average and median bias differed by approximately 5–15%, depending on the level of uncertainty, the ratio of the crude and adjusted risk estimates suggested that median bias estimate is a more appropriate summary measure of the central tendency of bias. This can be attributed to the fact that mean bias compared to median bias is vulnerable to extreme samples from the Dirichlet distribution in the sampling procedure. As such, confounding by smoking was estimated at about 13%, which is somewhat higher than when using the Axelson and Steenland method[9] used by Hooiveld *et al*[16] (range 9–12%, calculated from the manuscript), and suggests similar exposure–response associations as shown in the original publications.[21] Similarly, assuming a correlation between smoking habits and average bitumen fume exposure, which results from longer breaks during tasks and hence more opportunity for smoking breaks or from ignoring side effects from smoking given the working circumstances during the highest exposed tasks, shows comparable trends with reduced levels of certainty (wider credible intervals).

Table 1 Relative risks for lung cancer due to lifetime bitumen fume exposure (RR), 95% confidence interval (CI) and 95% credible intervals (CL), estimated from a frequentist Poisson regression model, a Bayesian Poisson model and an adjusted Bayesian Poisson model

	Frequentist model, RR_{unadj} (95% CI)	Bayesian model, RR_{unadj} (95% CL)	Bayesian adjusted model (5% variance), RR (95% CL)	Bayesian adjusted model (10% variance), RR (95% CL)
Average exposure (mg/m³)				
<0.97	1 (–)	1 (–)	1 (–)	1 (–)
0.97–1.24	1.87 (1.17 to 3.00)	1.93 (1.16 to 2.99)	1.78 (0.75 to 3.55)	1.88 (0.56 to 4.61)
1.24–1.39	2.35 (1.32 to 4.18)	2.45 (1.30 to 4.17)	2.26 (0.88 to 4.78)	2.40 (0.67 to 6.09)
>1.39	2.73 (1.56 to 4.78)	2.82 (1.53 to 4.73)	2.61 (1.02 to 5.44)	2.76 (0.79 to 6.95)
Mean bias			19.4% (−38.5% to 164.0%)	27.7% (−55.5% to 198.4%)
Median bias			13.0% (−38.5% to 164.0%)	13.1% (−55.5% to 198.4%)
Cumulative exposure (years-mg/m³)				
<0.76	1 (–)	1 (–)	1 (–)	1 (–)
0.76–2.98	1.10 (0.51–2.35)	1.25 (0.55 to 2.59)	1.13 (0.37 to 2.73)	1.20 (0.30 to 3.33)
2.98–9.03	1.27 (0.48–3.36)	1.54 (0.52 to 3.78)	1.39 (0.37 to 3.76)	1.48 (0.30 to 4.56)
>9.03	2.00 (0.68–5.87)	2.46 (0.71 to 6.46)	2.22 (0.52 to 6.33)	2.35 (0.42 to 7.61)
Mean bias			19.5% (−38.7% to 115.1%)	27.6% (−55.0% to 196.5%)
Median bias			13.1% (−38.7% to 115.1%)	13.2% (−55.0% to 196.5%)

Table 2 Relative risks (RR), 95% confidence interval (CI) and 95% credible intervals (CL), estimated from a frequentist Poisson regression model and the Bayesian Poisson model, and Bayesian adjusted relative risks for the 15-year lagged model

	Frequentist model, RR_{unadj} (95% CI)	Bayesian model, RR_{unadj} (95% CL)	Bayesian adjusted model (5% variance), RR (95% CL)	Bayesian adjusted model (10% variance), RR (95% CL)
15-year lagged average exposure (years-mg/m^3)				
Non-exposed	1 (–)	1 (–)	1 (–)	1 (–)
<1.11	0.38 (0.21 to 0.70)	0.39 (0.20 to 0.68)	0.36 (0.14 to 0.78)	0.39 (0.11 to 0.99)
1.11–1.32	0.72 (0.40 to 1.31)	0.75 (0.39 to 1.28)	0.69 (0.26 to 1.47)	0.73 (0.20 to 1.88)
1.32–1.48	1.23 (0.68 to 2.23)	1.27 (0.66 to 2.19)	1.18 (0.45 to 2.50)	1.25 (0.35 to 3.21)
>1.48	1.89 (0.96 to 3.70)	1.94 (0.91 to 3.54)	1.80 (0.63 to 3.96)	1.91 (0.50 to 5.06)
Mean bias			19.6% (−38.2% to 116.0%)	27.3% (−55.2% to 195.4%)
Median bias			13.1% (−38.2% to 116.0%)	13.0% (−55.2% to 195.4%)
15-year lagged cumulative exposure (years-mg/m^3)				
Non-exposed	1 (–)	1 (–)	1 (–)	1 (–)
<0.90	0.72 (0.36 to 1.44)	0.73 (0.33 to 1.37)	0.68 (0.23 to 1.52)	0.72 (0.18 to 1.91)
0.90–2.83	0.82 (0.48 to 1.41)	0.84 (0.47 to 1.39)	0.77 (0.31 to 1.60)	0.82 (0.24 to 2.05)
2.83–8.96	0.74 (0.41 to 1.33)	0.76 (0.40 to 1.31)	0.70 (0.27 to 1.50)	0.74 (0.21 to 1.91)
>8.96	1.37 (0.61 to 3.10)	1.45 (0.58 to 3.00)	1.34 (0.41 to 3.24)	1.42 (0.33 to 3.99)
Mean bias			19.4% (−38.5% to 158.0)	27.8% (−55.5% to 198.4%)
Median bias			12.9% (−38.5% to 158.0%)	13.5% (−55.5% to 198.4%)

The estimation of the bias factor depended on the assumptions regarding prevalence of smokers, ex-smokers and never smokers in the exposed and referent populations in the cohort, which were extrapolated from a Dutch sub-cohort and need not necessarily be valid for the complete multi-centre cohort. This has been reflected in a priori selecting two different levels of variance for the Dirichlet distribution. However, smoking data were also collected specifically for bitumen exposed and non-exposed workers in the Finnish sub-cohort,[28] and showed that the percentages of ever and never smokers in the bitumen exposed workers (77% and 23%, respectively) and non-exposed workers (68% and 32%) were similar to those in the Netherlands (79% and 21% for exposed and 72% and 28% for non-exposed, respectively). As such, a 5% variance in priors on prevalence of smoking habits might be more appropriate than a 10% variance.

The use of a Dirichlet distribution to estimate the prevalence of current, former and ex-smokers has the advantage over a multivariate normal distribution that scenarios with a negative probability do not occur, and hence no posterior assessment of values and subsequent discarding of scenarios with negative probabilities is needed.[6] However, the variability in estimated bias in different scenarios is highly dependent on the variance of the probabilities, which were defined a priori in these simulations. As such, there is a large difference between the estimated mean of the bias and the median, which in our example leads to unrealistic mean bias factors of 92% (median is 12%) when variance is a priori set to 30% (data not shown). We argue that that such uninformative prior distributions should be avoided, since the usefulness of this method is based on improving the risk estimates based on data from other sources as a basis for the prior distributions. Nonetheless, reporting the median bias factor instead of the mean in future studies seems more appropriate since the median is less sensitive to outliers.

Finally, this method assumes that the unmeasured confounder (ie, smoking status) is not associated with the observed confounders (country, age, year, duration of employment and use of coal tar) in the model. This is a common assumption in similar Bayesian work.[29]

Despite these limitations, taken together with sensitivity analyses to assess the method of exposure assessment and lagging of exposure, as well as residual confounding from coal tar use in the original cohort,[30] our current results suggest that the exposure–response association in the original analyses could

Table 3. Relative risks (RR) and 95% credible intervals (CL) from the adjusted Bayesian Poisson model, assuming a positive association between average bitumen fume exposure and smoking prevalence

Average exposure (mg/m^3)	Bias (mean)	Bias (median)	95% CL	Bayesian adjusted model (5% variance), RR (95% CL)
<1.04	–			1
1.04–1.24	9.6%	3.5%	(−45.9% to 99.8%)	1.96 (0.80 to 4.00)
1.24–1.39	19.6%	13.1%	(−38.4% to 116.0%)	2.26 (0.86 to 4.78)
>1.39	29.2%	22.1%	(−32.0% to 132.3%)	2.39 (0.94 to 4.94)
Average exposure (mg/m^3)	**Bias (mean)**	**Bias (median)**	**95% CL**	**Bayesian adjusted model (10% variance), RR (95% CL)**
<1.04	–			1
1.04–1.24	9.6%	3.3%	(−58.5% to 116.4%)	2.09 (0.75 to 5.01)
1.24–1.39	19.6%	12.9%	(−51.5% to 131.3%)	2.39 (0.83 to 5.70)
>1.39	29.3%	22.7%	(−45.9% to 145.1%)	2.52 (0.91 to 5.89)

Main messages

- Applying the full Bayesian adjustment method to adjust for an unmeasured confounder proposed in this manuscript can be used to quantitatively evaluate the combined impact of an unmeasured confounding factor and sampling variability (typically reflected in frequentist confidence intervals); currently popular methods for adjustment for latent confounders in occupational epidemiology fail to reflect sampling variability and thus do not correctly integrate all sources of error.
- The analytical frameworks for Bayesian sensitivity analysis that we propose can be extended to a variety of epidemiological applications and would yield a more realistic appraisal of uncertainties in epidemiological results.
- Confounding from smoking in a large European multi-centre study of asphalt workers and lung cancer risk does not appear to fully explain the exposure–response association between bitumen fume exposure and increased lung cancer risk.

Policy implications

- Our results lend further support to the hypothesis that bitumen fume in the asphalt industry is an occupational carcinogen.
- Our method establishes a framework for (a) realistic representation of uncertainty in epidemiological results combined with (b) Bayesian estimation of the distribution of the true exposure–response gradient, both of which are essential for rigorous risk assessment that should inform rational policy towards control of hazardous exposures.

not be completely attributed to confounding by smoking or other confounding factors. Although Mundt and colleagues[17] also used a method similar to that described here to adjust the results of the asphalt cohort for confounding by smoking, the fact that their conclusions differ from ours can be attributed to different specification of the prior distributions. Their confounded risk estimates were mainly based on cohort studies comparing lung cancer rates in asphalt workers/roofers with smoking rates among construction workers, which both have a higher smoking prevalence than the general population,[31] and they used smoking prevalence rates in the general population obtained from the United States National Health Interview Survey (NHIS)[32] as a comparison. Thus, although half of the cohort studies Mundt *et al* examined were conducted in Europe, they used smoking rates for the US that are generally considered to be lower than those in Europe. As such, these assumptions are likely to inflate the bias factor. Similarly, a meta-analysis by Fayerweather[33] addressing residual confounding by coal tar use also depended on the unrealistic assumptions about the exposure distributions as discussed in detail by Burstyn and Kromhout.[34] This indicates that great care must be taken in defending assumptions used in sensitivity analysis, but it must be said in defence of quantitative sensitivity analyses that the transparent nature of assumptions they entail elevates the discussion of latent confounding from the realm of speculation to numerical analysis.

In summary, we describe the application of a fully Bayesian adjustment method to account for an unmeasured confounding factor and its uncertainty, as well as the uncertainty due to random sampling error in the disease model. Applying this method to the question of residual confounding by smoking in a European asphalt workers' cohort study assessing average, cumulative and 15-year lagged exposure to bitumen fume and increased lung cancer risk, demonstrated that the bias arising from smoking habits is relatively small and is unlikely to explain the association between increased lung cancer risk and career average bitumen fume exposure found in the original study. However, external adjustments, even with the most defensible assumptions about the relevant distributions, always require a leap of faith and therefore cannot replace internal adjustments by actual assessment of the smoking history for the cohort participants.

Acknowledgements: . The authors would like to thank Dr Paul Gustafson for statistical advice.

Funding: The Asphalt Workers' cohort study was sponsored by the European Commission (grant number: BMH4-CT95-1100), EAPA, Eurobitume and CONCAWE. The project described in this manuscript was conducted within a training fellowship grant from the European Union Sixth Framework Programme Network of Excellence on Environmental Cancer Risk, Nutrition and Individual Susceptibility (ECNIS) (grant number: FOOD-CT-2005-513 943).

Competing interests: None.

REFERENCES

1. **Blair A,** Stewart P, Lubin JH, *et al*. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *Am J Ind Med* 2007;**50**:199–207.
2. **Grimes AD,** Schulz KF. Bias and causal associations in observational research. *Lancet* 2002;**359**:248–52.
3. **Ahlbom A,** Steineck G. Aspects of misclassification of confounding factors. *Am J Ind Med* 1992;**21**:107–12.
4. **Fewell Z,** Davey SG, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol* 2007;**166**:646–55.
5. **Greenland S,** Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol* 1986;**15**:413–19.
6. **Steenland K,** Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am J Epidemiol* 2004;**160**:384–92.
7. **Sturmer T,** Schneeweiss S, Avorn J, *et al*. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005;**162**:279–89.
8. **Greenland S.** Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996;**25**:1107–16.
9. **Axelson O,** Steenland K. Indirect methods of assessing the effects of tobacco use in occupational studies. *Am J Ind Med* 1988;**13**:105–18.
10. **Kriebel D,** Zeka A, Eisen EA, *et al*. Quantitative evaluation of the effects of uncontrolled confounding by alcohol and tobacco in occupational cancer studies. *Int J Epidemiol* 2004;**33**:1040–5.
11. **MacLehose RF,** Kaufman S, Kaufman JS, *et al*. Bounding causal effects under uncontrolled confounding using counterfactuals. *Epidemiology* 2005;**16**:548–55.
12. **Greenland S.** Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal* 2001;**21**:579–83.
13. **Molitor J,** Molitor NT, Jerrett M, *et al*. Bayesian modeling of air pollution health effects with missing exposure data. *Am J Epidemiol* 2006;**164**:69–76.
14. **Gryparis A,** Coull BA, Schwartz J. Controlling for confounding in the presence of measurement error in hierarchical models: a Bayesian approach. *J Expo Sci Environ Epidemiol* 2007;**17**:S20–S28.
15. **Binet S,** Pfohl-Leszkowicz A, Brandt H, *et al*. Bitumen fumes: review of work on the potential risk to workers and the present knowledge on its origin. *Sci Total Environ* 2002;**300**:37–49.
16. **Hooiveld M,** Spee T, Burstyn I, *et al*. Lung cancer mortality in a Dutch cohort of asphalt workers: evaluation of possible confounding by smoking. *Am J Ind Med* 2003;**43**:79–87.
17. **Mundt DJ,** van Wijngaarden E, Mundt KA. An assessment of the possible extent of confounding in epidemiological studies of lung cancer risk among roofers. *J Occup Environ Hyg* 2007;**4**(S1):163–74.
18. **Schulte PA.** Gaps in scientific knowledge about the carcinogenic potential of asphalt/bitumen fumes. *J Occup Environ Hyg* 2007;**4**(Suppl 1):3–5.
19. **Partanen T,** Boffetta P. Cancer risk in asphalt workers and roofers: review and meta-analysis of epidemiologic studies. *Am J Ind Med* 1994;**26**:721–40.
20. **Boffetta P,** Burstyn I, Partanen T, *et al*. Cancer mortality among European asphalt workers: an international epidemiological study. I. Results of the analysis based on job titles. *Am J Ind Med* 2003;**43**:18–27.

21. **Boffetta P,** Burstyn I, Partanen T, *et al*. Cancer mortality among European asphalt workers: an international epidemiological study. II. Exposure to bitumen fume and other agents. *Am J Ind Med* 2003;**43**:28–39.
22. **IARC**. *Tobacco smoke and involuntary smoking*. IARC monographs on the evaluation of carcinogenic risks to humans. Vol 83. Lyon: IARC Press, 2004.
23. **Burstyn I,** Boffetta P, Kauppinen T, *et al*. Performance of different exposure assessment approaches in a study of bitumen fume exposure and lung cancer mortality. *Am J Ind Med* 2003;**43**:40–8.
24. **Burstyn I,** Boffetta P, Kauppinen T, *et al*. Estimating exposures in the asphalt industry for an international epidemiological cohort study of cancer risk. *Am J Ind Med* 2003;**43**:3–17.
25. **Simonato L,** Agudo A, Ahrens W, *et al*. Lung cancer and cigarette smoking in Europe: an update of risk estimates and an assessment of inter-country heterogeneity. *Int J Cancer* 2001;**91**:876–87.
26. **Connor RJ,** Mosimann JE. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J Am Stat Assoc* 1969;**64**:194–206.
27. **Lunn DJ,** Thomas A, Best N, *et al*. WINBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000;**10**:325–37.
28. **Kauppinen T,** Heikkila P, Partanen T, *et al*. Mortality and cancer incidence of workers in Finnish road paving companies. *Am J Ind Med* 2003;**43**:49–57.
29. **McCandless LC,** Gustafson P, Levy A. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat Med* 2007;**26**:2331–47.
30. **de Vocht F,** Burstyn I, Ferro G, *et al*. Sensitivity of the association between increased lung cancer risk and bitumen fume exposure to the assumptions in the assessment of exposure. *Int Arch Occup Environ Health* 2008 Oct 28 [Epub ahead of print].
31. **Sorensen G,** Barbeau E, Hunt MK, *et al*. Reducing social disparities in tobacco use: a social-contextual model for reducing tobacco use among blue-collar workers. *Am J Public Health* 2004;**94**:230–9.
32. **Lee DJ,** LeBlanc W, Fleming LE, *et al*. Trends in US smoking rates in occupational groups: the National Health Interview Survey 1987–1994. *J Occup Environ Med* 2004;**46**:538–48.
33. **Fayerweather WE.** Meta-analysis of lung cancer in asphalt roofing and paving workers with external adjustment for confounding by coal tar. *J Occup Environ Hyg* 2007;**4**:175–200.
34. **Burstyn I,** Kromhout H. Still no evidence that coal tar exposure confounded the association between bitumen/asphalt fume and lung cancer in the cohort of European asphalt workers. *J Occup Environ Hyg* 2008;**5**:D73–D75.

APPENDIX A

```
# Specification of the Bayesian Poisson model with external adjustment for confounder (smoking)
Model {
for(i in 1:N){
y[i]~ dpois(mu[i])
log(mu[i])<- int+offset[i]+g1[i]+g2[i]+g3[i]+g4[i] +g5[i]
g1[i]<- b1*exph[i]+b2*expm[i]+b3*expl[i]+b4*age10[i]+b5*age9[i]
g2[i]<-b6*age8[i]+b7*age7[i]+b8*age6[i]+b9*age5[i]+b10*age4[i]
g3[i]<- b11*age3[i]+b12*age2[i] +b13*cal5[i] +b14*cal4[i] +b15*cal3[i]
g4[i]<- b16*cal2[i]+b17*alljc7[i] +b18*alljc6[i] +b19*alljc5[i] +b20*alljc4[i]
g5[i]<- b21*c8[i] +b22*c7[i] +b23*c6[i] +b24*c5[i] +b25*c4[i] +b26*c2[i] +b27*coaltar[i]
}
# Specification of the prior distributions
int~dnorm(0,0.0001)
b1~dnorm(0,0.0001)
…
b27~dnorm(0,0.0001)
betacur~dnorm(3.17,10.14)
betaform~dnorm(2.01,10.30)
# Estimation of tobacco smoking prevalence in the exposed strata
psmkexp[1:3]~ddirch(alphaexp[])
psmknonexp[1:3] ~ ddirch(alphanonexp[])
pnev1<-psmkexp[1]
pnev0<-psmknonexp[1]
pcur1<-psmkexp[3]
pcur0<-psmknonexp[3]
pform1<-psmkexp[2]
pform0<-psmknonexp[2]
# Estimation of variance of tobacco smoking prevalence in the exposed strata
a0_exp<-pnev1+pcur1+pform1
a0_nonexp<-pnev0+pcur0+pform0
var_pnev1<-(pnev1*(a0_exp-pnev1))/((a0_exp*a0_exp)*(a0_exp+1))
var_pnev0<-(pnev0*(a0_nonexp-pnev0))/((a0_nonexp*a0_nonexp)*(a0_nonexp+1))
var_pcur1<-(pcur1*(a0_exp-pcur1))/((a0_exp*a0_exp)*(a0_exp+1))
var_pcur0<-(pcur0*(a0_nonexp-pcur0))/((a0_nonexp*a0_nonexp)*(a0_nonexp+1))
var_pform1<-(pform1*(a0_exp-pform1))/((a0_exp*a0_exp)*(a0_exp+1))
var_pform0<-(pform0*(a0_nonexp-pform0))/((a0_nonexp*a0_nonexp)*(a0_nonexp+1))
# Estimation of unknown bias factor
bias<-(pnev1+exp(betacur)*(pcur1)+exp(betaform)*(pform1))/(pnev0+exp(betacur)*(pcur0)+exp(betaform)*(pform0))
# Estimate of Relative risks uncorrected for confounding by smoking
RR.high<-exp(b1)
RR.med<-exp(b2)
RR.low<-exp(b3)
# Estimate of Relative risk adjusted for smoking
RR.high.adj<-RR.high/bias
RR.med.adj<-RR.med/bias
RR.low.adj<-RR.low/bias
}
#specification of initial MCMC starting values based on means, 2.5%Confidence Limit, and 97.5%Confidence Limit of the frequentist poisson model
#data
#set-up Dirchelet prior: illustrated values are for 10% variance assuming no
# increase in prevalence of smoking with increase in exposure
# (tables 1 and 2 in text)
list(alphaexp = c(1.113, 1.696, 2.491), alphanonexp = c(1.568,1.792,2.240), …
# where:
# N = number of observations in the cohort
# int = model intercept (α)
# betacur = model parameter estimate (β) for current tobacco smokers
# betaform = model parameter estimate (β) for former tobacco smokers
# pnev = proportion of never smokers among the exposed (1) and unexposed (0)
# pcur = proportion of current smokers among the exposed (1) and unexposed (0)
# pform = proportion of former smokers among the exposed (1) and unexposed (0)
```