BAYESIAN METHODS TO ADDRESS MULTIPLE COMPARISONS AND MISCLASSIFICATION BIAS IN STUDIES OF OCCUPATIONAL AND ENVIRONMENTAL RISKS OF CANCER

A thesis by publications presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

Massey University, Wellington, New Zealand

Marine Corbin

2013

Abstract

In this thesis I explore the application of several Bayesian approaches, implemented with standard statistical software, in environmental and occupational epidemiology. These methods are applied to case-control studies of occupational risks for lung and upper aerodigestive tract cancers conducted in New Zealand and Europe. The findings are of interest in themselves, but the focus of the thesis is on the application of Bayesian methods to produce these findings. It is not intended to represent a comprehensive overview of all Bayesian methods, but rather to explore Bayesian methods which are most appropriate for the studies which are presented here.

In the first section, I review the underlying theory involved in such analyses.

In the second section, I use Bayesian methods to address the problem of multiple comparisons. In occupational case-control studies, we may collect information on hundreds of occupations/exposures for which there is little or no prior evidence. For those occupations/exposures, we get a false positive finding by chance about 5% of the time. This means that if we repeat the study in a new population, these chance associations are likely to exhibit 'regression to the mean' and will not show such extreme risks again. Bayesian methods can be used to 'shrink' effect estimates based on how strong the regression to the mean is likely to be.

In the third section, I use Bayesian methods for assessing and correcting systematic error. Although the methods I use can be applied to several situations (selection bias, misclassification, residual confounding), I apply them to the specific situation of

misclassification of the main exposure. In particular, I apply four different methods for such sensitivity analyses: multiple imputation for measurement error (MIME); imputation based on specifying the sensitivity and specificity (SS), Direct Imputation (DI) of the 'true' exposure using a regression model for the predictive values and imputation based on a fully Bayesian analysis.

I conclude by summarising the strengths, limitations, and areas of future development for the use of these methods. It is anticipated that, in 5-10 years time, such analyses may become standard supplements to 'traditional' forms of analysis, i.e. that Bayesian methods may be routinely used, and may form part of the 'epidemiological toolkit' for assessing and correcting for both random and systematic error.

Author's declaration

This thesis was produced according to Massey University's "Thesis-by-Paper" requirements. That is, it is based on research that is published, in-press, submitted for publication, or is in final preparation for submission. Each individual chapter is set out in the style of the journal to which it has been submitted. Consequently, some of the submitted chapters are relatively succinct, there is some repetition (particularly in the Methods sections) and there are small stylistic differences between chapters. To supplement the relative brevity of some of the chapters, the appropriate sections of the background and methods chapter have been extended.

I have stated my contribution to each chapter in Appendix IV.

Acknowledgements

First of all I would like to thank both of my supervisors Neil Pearce and Milena Maule for their constant guidance, support and faith in me during this long adventure and through the distance.

Neil, thank you for welcoming me in New Zealand and at Centre for Public Health Research (CPHR) and for giving me the chance to embark on this PhD. Thanks for your advice and encouragement over the years and thanks for all the opportunities you gave me to extend my knowledge and experience. Thanks also for always finding the time and the ways to meet regularly and answer my questions, even though the different locations, internet connection and time differences did not always make it very easy.

Milena, grazie della tua amicizia e di essere sempre stata qua per me durante tutti questi anni, anche durante i primi mesi di vita di Matteo. Grazie di avermi dato la motivazione e di avermi incoraggiata a iniziare questo dottorato. Grazie del tuo immenso aiuto sia sul piano lavorativo che sul piano morale e di aver condiviso con me tutti i momenti alti e bassi nella realizzazione di questa tesi. Sei stata bravissima a saper ridarmi energia e fiducia ogni volta che ne avevo bisogno e ricorderò sempre sia le insalate di formule che tutte le risate insieme.

Thanks to all my workmates at CPHR for making me feel so quickly part of the 'family'. In particular, thanks to Jeroen Douwes for welcoming me back at CPHR for the last part of my PhD and for helping me through the examination process. Thanks to my mock examiners Jeroen, Steve Haslett, Laura Howe, Andrea 't Mannetje, Amanda Eng and Collin Brookes for their constructive comments. Thanks to Steve for his availability and for his very helpful guidance. Thanks to Amanda and Collin, my

“mentor PhD students”, for all their valuable advice and support. Thanks to Dave McLean, Andrea ‘t Mannetje, Soo Cheng and Fiona McKenzie for their help with the lung cancer study. Thanks to Mathu and Helene for their support and coaching and for our weekly quiz nights and a particular thank you Mathu for hosting me in your lovely apartment every time I came back to Wellington. Thanks to Katharine for being such a supportive roommate during the ultimate phase of this PhD. Thanks to Soo and Grace for keeping me going with the magic tiger balm and essential oils. Thanks to Kerry and Soo for the many rides home when I stayed late at work. Thanks also to Hilary for being always so helpful and thanks to Nathalie and Vicki for their help in the last minute rush.

Grazie ai miei colleghi dell’Unità di Epidemiologia dei Tumori per la loro accoglienza e per avermi viziata dal mio primo giorno a Torino. Innanzitutto grazie mille a Franco Merletti di avermi accolta prima come stagista e poi come dottoranda, di avermi spinta e indirizzata nella scelta di questo dottorato e di avermi dato tutte le opportunità possibili per condurre questo progetto. Grazie a Lorenzo Richiardi per il suo importante contributo a questa tesi e per i suoi consigli che mi hanno aiutata tante volte. Grazie a tutti i “stanzonesi” (Milena, Lorenzo, Daniela, Costanza, Emanuele e Enrica) per tutti i buoni momenti passati insieme e i tradizionali pranzi dagli “Oscar” che mi mancano. Grazie anche a Daniela Aimar di avermi ospitata durante alcune settimane nella sua casa.

I want to thank all my coauthors and in particular thanks to Sander Greenland and Kyle Steenland for their guidance and advice. Thanks also to Jonathan Bartlett for his help and input on Chapter VI.

My stay in New Zealand would not have been such a nice experience if I had not had a nice home to go to every night. Thanks to Carl Lin, Jacob, KC, Steve Mainwaring, Mousumi, Matilda and Swann for being such amazing flatmates.

I am grateful to all my friends for always staying in contact even through the distance. Un spécial gros merci à Claire, Morgane, Elena et Manue pour leurs visites à Wellington et/ou à Turin qui m'ont fait énormément plaisir.

I also wish to thank my wonderful family. Merci à Maman, Martin, Clémentine, Marjolaine, Corentin et Capucine de m'avoir soutenue et encouragée pendant toutes mes études. Merci d'avoir supporté mes crises de nerfs à chaque départ, quand je décidais de déballer ma valise cinq minutes avant de partir parce que j'avais oublié quelque chose. Merci aussi d'avoir fait le voyage tous les six à Turin et à Welllington. Merci à Papy et Mamie de m'avoir également soutenue et accompagnée dans tous mes projets. Merci de m'avoir emménagée et déménagée lors de tous mes déplacements en Europe (même sous la neige) et de m'avoir continuellement gâtée. Merci aussi de vos 2 consécutives visites en Nouvelle-Zélande ! Merci aussi à ma cousinette Hélène de m'avoir hébergée lors de mes visites à Londres.

Finally, Sebastián, without this PhD I would probably have never met you but without you I would probably have never managed to finish this thesis in time. Thanks for your support for the last few years and for always believing in me and thanks also for your very special care for the last months. ¡Mil gracias por todo!

Table of Contents

List of tables

List of figures

List of Abbreviations

CI	Confidence interval
CL	Confidence/Credibility limits
DI	Direct imputation of the ‘true’ exposure using a regression model for the predictive values
EB	Empirical Bayes
FBA	Fixed-parameter bias analysis
HR	Hierarchical regression
ISCO	International Standard Classification of Occupations
ISIC	International Standard Industrial Classification
logOR (or ln(OR))	log Odds Ratio
MCMC	Markov Chain Monte Carlo
MIME	Multiple imputation for measurement error
ML	Maximum likelihood
NACE	National Industrial Classification of All Economic Activities
NZSCO	New Zealand Standard Classification of Occupations
NZSEI	New Zealand Socio-Economic Index
OR	Odds Ratio
PBA	Probabilistic bias analysis
SB	Semi-Bayes
SI	Simulation Intervals
SL	Simulation Limits
SS	Imputation based on specifying the sensitivity and specificity
UADT	Upper aerodigestive tract

SECTION 1

Introduction, background and methods

CHAPTER I

General introduction

Bayesian thinking is part of our daily life. We almost always base decisions on past experience, together with any new information that we may have[1]. We expect the sun to rise tomorrow much as it did yesterday and today. Whenever we hear the news, we interpret it in light of what we know already. We don't start each day with a clean slate and an empty mind.

Bayesian methods were used alongside other statistical methods from the eighteenth century until the 1920s[2]. At that time, several influential statisticians (Fisher, Neyman, Pearson) developed frequentist techniques. These analytical methods were developed and initially applied mainly in the context of large randomised controlled trials and for large agricultural experiments, and they generally work well in those situations. They assume that the trial data have been sampled from an infinite population (i.e. that the trial is a sample of an infinite number of possible trials of the same size), and that exposure has been randomised. So, for example, the idea behind the estimation of a 95% confidence interval is that, if we repeated the study an infinite number of times, and if exposure had been randomised, there was no confounding or misclassification and perfect response rates, then the estimated confidence interval would include the true effect measure 95% of the time.

In epidemiological studies, both exposures and outcomes are observed. Therefore, the validity of the results obtained with frequentist techniques becomes questionable as these studies cannot be repeated an infinite number of times and do not involve randomised exposures, thus making confounding very likely. In addition, response rates are often low and there are serious problems of misclassification. Further uncertainty is introduced when deciding what variables to include in our models, how to categorize them, what form the dose–response should take for a continuous variable, what to do about interactions, and how to assess confounding.

Most epidemiologists operate as Bayesians. They decide what to study based on prior knowledge. For example, one might study pesticides and non-Hodgkin lymphoma because many studies have shown an increased risk of this lymphoma in farmers, and some studies have specifically linked the increased risk to pesticide use. This prior evidence will usually be summarized in the grant application and in the Introduction section of the report from the subsequent study. The findings are usually also interpreted in a Bayesian framework in the Discussion section of the paper, i.e. in light of what the study adds to prior knowledge. The Discussion section often considers, for example, the results of previous related studies and considers the biological plausibility of the findings. However, the statistical analyses are usually carried out in a frequentist framework, as if the specific study data are the only data which exist.

In contrast with frequentist approaches, Bayesian statistical approaches formally take prior evidence into account in the analysis. Thus, they consider prior evidence in the Methods and Results sections, as well as in the Introduction and Discussion. Unfortunately, these methods have become associated with highly complex approaches,

such as Markov chain Monte Carlo methods, while, in fact, most Bayesian methods can be applied using ordinary statistical software packages, or with a simple spreadsheet or hand calculator[2].

The main objective of this thesis is to explore the application of several Bayesian approaches, implemented with standard statistical software, in environmental and occupational epidemiology. These methods are applied to several different studies of occupational risks for lung cancer and upper aerodigestive tract cancer, and the findings are therefore of interest in themselves. However, the focus of the thesis is on the application of Bayesian methods to produce these findings.

Section 1. Introduction, background and methods

Chapter II: Background and methods

Chapter II presents the context of this thesis and gives a general overview of the methods applied in the following chapters. In the first part of the chapter, I discuss studies of occupational and environmental cancer, including the methodological issues of random error (chance) and systematic error (bias). In the second half, after a brief introduction to Bayesian theory, I outline Bayesian approaches for addressing random and systematic error. These methods are then applied in the following chapters.

Section 2. Random error

Chapter III, Chapter IV and Chapter V explore the use of Bayesian methods for addressing random error, and in particular, random error arising from the estimation of multiple associations in occupational cancer studies.

Chapter III: Lung cancer and occupation: A New Zealand cancer registry-based case–control study

Chapter III presents the results of an occupational case-control study of lung cancer in New Zealand[3]. Risks of lung cancer were estimated separately for each occupation, adjusting for gender, age, ethnicity, smoking, and socio-economic status and a Semi-Bayes approach was applied to adjust for multiple comparisons.

Chapter IV: Occupation and risk of upper aerodigestive tract cancer: the ARCAGE study

Chapter IV reports the association between upper aerodigestive tract (UADT) cancer and occupational history in the European case-control study ARCAGE[4]. Risks of UADT cancer were estimated distinctly in men and women for each occupation and for each industry, adjusting for age, study centre, smoking and alcohol. A Semi-Bayes approach was applied to adjust for multiple comparisons.

Chapter V: Hierarchical regression for multiple comparisons in a case-control study of occupational risks for lung cancer

Chapter V goes beyond the work presented in chapters III and IV by adopting a hierarchical (multi-level) approach. It involves an application of hierarchical regression to adjust for multiple comparisons in a study of occupational risk factors for lung cancer in Italy[5]. Risks of lung cancer were estimated separately for each occupation and a second-stage model was added to include prior information on exposures to three lung carcinogens (asbestos, chromium and silica) for each occupation. The results obtained with hierarchical regression were then compared with those obtained with 'non-hierarchical' Semi-Bayes adjustment towards the global mean.

Section 3. Systematic error

Chapter VI: Adjustment for exposure misclassification – Application of several methods in a case-control study of lung cancer where the smoking status has been misclassified

Chapter VI explores methods addressing systematic error, and in particular, misclassification bias. This chapter explores four different approaches to adjust for exposure misclassification while estimating an exposure-disease association in the presence of other covariates: multiple imputation for measurement error (MIME); and three types of sensitivity analysis: imputation based on specifying the sensitivity and specificity (SS), Direct Imputation (DI) of the 'true' exposure using a regression model for the predictive values, and imputation based on a fully Bayesian analysis[6]. These four methods are applied to estimate the association between smoking status and lung cancer in simulated data obtained by misclassifying the smoking status data from the study analysed in Chapter III.

Section 4. Discussion and conclusions

Chapter VII: General discussion

Finally, Chapter VII summarises the main findings of the thesis. The limitations of the methods are discussed and recommendations are made for future research and applications of Bayesian methodology in occupational and environmental epidemiology. I also summarize the substantive findings of the papers included in this thesis and give recommendations for future research into occupational and environmental causes of lung cancer and upper aerodigestive tract cancer.

CHAPTER II

Background and methods

In this chapter I discuss the background and general methodological issues which form the basis for the studies which are presented in the following chapters. I do not attempt to give a comprehensive overview of occupational cancer research, or of Bayesian methods. Rather, I highlight the key issues which were relevant to the conduct of the studies which appear in the following chapters.

A. Background

1. Occupational and environmental risk factors for cancer

Carcinogenesis is a complex process. Moolgavkar's mathematical model (1978) on cancer causation states that carcinogenesis is composed of several different phases[7]: In the first phase called initiation, a healthy stem cell mutates, becoming an intermediate cell. Then, if an exposure enhances the multiplication of intermediate cells (promotion), it is more likely that one of them will grow into a malignant cell (progression). These changes may involve random genetic mutations, mutations caused by environmental exposures, or other effects of environmental exposures. For example, an exposure which causes multiplication of intermediate cells may increase the risk of cancer even though

it does not cause mutations. Thus, cancer occurs because of a complex interplay between both genetic and environmental factors.

Environmental factors may be particularly important in terms of explaining changes over time, and differences between populations. For example, Japanese migrating to the United States experience increased risks of colon and breast cancer and decreased risks of stomach cancer[8-10]. Another indication of the effect of environmental factors on cancer is the variation of cancer rates over time such as the steep increase in lung cancer incidence in the second half of the 20th century following increases in cigarette consumption in the first half of the century[11].

The first clear-cut evidence of an association between cancer and an occupational exposure was found in 1775 by Percival Pott, who discovered that soot was responsible for cancer in London chimney sweeps[12]. An important number of occupational exposures have subsequently been shown to be associated with higher risks of cancer. The International Agency for Research on Cancer (IARC) has determined a classification of specific agents according to their evidence of carcinogenicity in the IARC monographs[13]. The main environmental factors which have been established human carcinogens in this classification include air pollutants (such as asbestos) and water pollutants (such as arsenic), radiation and in particular ultraviolet radiation, viruses (such as hepatitis B and C and Human Papilloma Virus (HPV)), aflatoxins, drugs and chemical agents occurring in occupational settings.

Table II.1 shows the 22 agents, for which exposures are mostly occupational, without considering pesticides and drugs, which are established human carcinogens (Group 1)[7, 13]. A further 20 agents have been classified as probably carcinogenic to humans (Group 2A) and other 20 as possible human carcinogens (Group 2B). Many associations have also been observed between cancer and particular occupations or industries but it is difficult to identify the causal exposures.

Environmental and occupational epidemiological studies attempt to estimate the risk of cancer associated with particular exposures or occupations in order to be able to prevent the disease through appropriate interventions. Occupational causes of cancer are particularly important in this regard, since prevention may be relatively easier (e.g. through health and safety regulations) than it is for 'lifestyle' exposures in the general population[14].

Table II.1. The 22 agents, for which exposures are mostly occupational (without considering pesticides and drugs) which are established human carcinogens (Group 1)[7, 13]

Exposure	Human target organ(s)	Main industry/use
4-Aminobiphenyl	Bladder	Rubber manufacture
Arsenic and arsenic compounds	Lung, skin	Glass, metals, pesticides
Asbestos	Lung, pleura, peritoneum	Insulation, filter material, textiles
Benzene	Leukaemia	Solvent, fuel
Benzidine	Bladder	Dye/pigment manufacture, laboratory agent
Beryllium and beryllium compounds	Lung	Aerospace industry/metals
Bis(chloromethyl)ether	Lung	Chemical intermediate/by-product
Chloromethyl methylether (technical grade)	Lung	Chemical intermediate/by-product
Cadmium and cadmium compounds	Lung	Dye/pigment manufacture
Chromium (VI) compounds	Nasal cavity, lung	Metal plating, dye/pigment manufacture
Coal-tar pitches	Skin, lung, bladder	Building material, electrodes
Coal-tars	Skin, lung	Fuel
Ethylene oxide	Leukaemia	Chemical intermediate, sterilant
Mineral oils, untreated and mildly treated	Skin	Lubricants
Mustard gas (sulphur mustard)	Pharynx, lung	War gas
2-Naphthylamine	Bladder	Dye/pigment manufacture
Nickel compounds	Nasal cavity, lung	Metallurgy, alloys, catalyst
Shale-oils	Skin	Lubricants, fuels
Soots	Skin, lung	Pigments
Talc containing asbestiform fibers	Lung	Paper, paints
Vinyl chloride	Liver, lung, blood vessels	Plastics, monomer
Wood dust	Nasal cavity	Wood industry

2. Statistical issues in the estimation of risks associated with occupational and environmental exposures

Epidemiological studies can produce inaccurate effect estimates due to two particular issues: random error (chance) and systematic error (bias).

Random error

Random error is an error which can be reduced towards zero when the size of the study is increased infinitely[15]. It is mainly due to the sampling variability, i.e. to the selection of the study participants, which are usually only a sample of the population.

In observational epidemiological studies, random error is present, together with the potential for systematic error. Random error becomes a particular issue in epidemiological studies that involve large numbers of comparisons. For example, environmental and occupational studies are often affected by the problem of multiple comparisons. This problem occurs because, in a given study where multiple measures are tested, as the number of tests increases, it becomes increasing likely to find statistically significant outcomes (false-positives, FP) due to random error, even if no real effects exist[16]. If the null hypothesis of a test is true, i.e. exposure actually has no effect on the risk of the disease, a statistically significant difference may still be observed by chance. This is termed type I error. For one test, the probability of having a type I error is usually set at given number α, which is termed the significance level. For n tests, the probability of having at least one type I error will be $1-(1-\alpha)^n$, and hence will increase as n increases. For instance, if we set α at 0.05, the probability of having at least one type I error for 10 tests will be 0.4, and for 100 tests will be 0.99[17].

Traditional methods of adjustment for multiple comparisons, such as the Bonferroni method[18] may induce investigators to ignore potentially important findings, because they do not take account of the fact that some variables are of greater a priori interest than others. Thus, by decreasing the probability of type I error, they increase the probability of type II error (the chance that carcinogenic exposures are not discovered)[19]. For example, consider a case-control study of asbestos exposure and lung cancer, in which a number of other exposures are also considered. The Bonferroni method involves 'adjusting' the significance level to take account of the number of comparisons involved, even though the a priori evidence is very strong for asbestos, but may be much weaker (or non-existent) for the other exposures being considered. Furthermore, the Bonferroni method only 'adjusts' estimates of statistical significance (p-values) and does not 'adjust' the effect estimates themselves (e.g. odds ratios and 95% CI) even though some of these may be biased away from the null due to random error.

Systematic error

Unlike random error, systematic error cannot be reduced by increasing the size of the study. Three main types of systematic error can be present in occupational and environmental studies: confounding, selection bias and information bias.

Confounding occurs when the effects of the exposure of interest on cancer risk are mixed with the effects of other factors[12]. If one or more confounders are not measured in the data (and are therefore not adjusted for), the exposure-disease effect estimate may be biased. Selection bias occurs when study participants are not chosen randomly from the source population, e.g. because of biased selection or because of low

response rates. Information bias occurs when study participants are misclassified with respect to their disease or their exposure status.

While it is possible to reduce bias in the study's design, confounding, selection bias and information bias should still be addressed in the statistical analysis. Most of the time, however, bias is only addressed qualitatively in the discussion section of a paper and assumed to be nonexistent in the data analysis[20].

B. Methods

1. Introduction to Bayesian inference

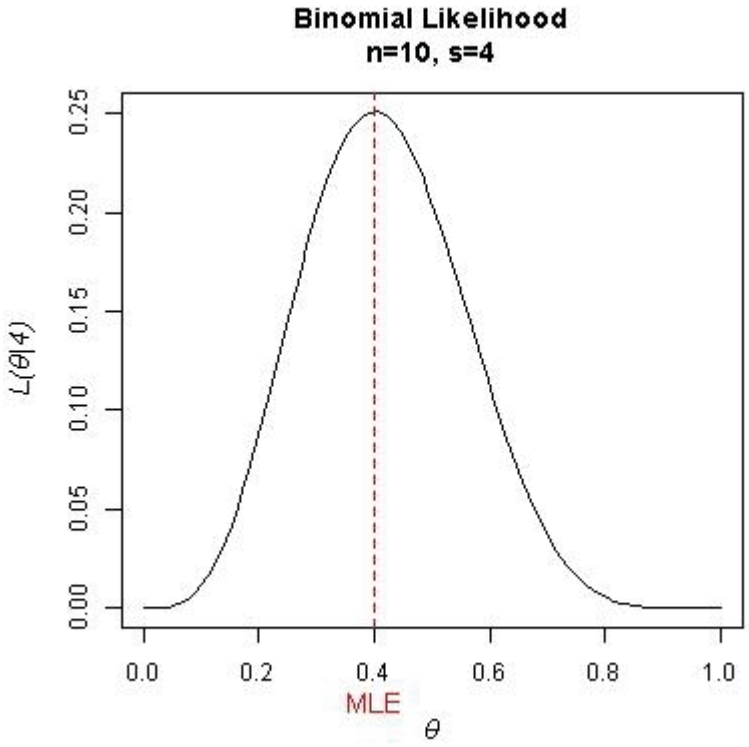
a) The likelihood function and the likelihood principle

Likelihood is a key notion in statistical inference, which is common to both the frequentist and Bayesian schools. The concept was introduced by Fisher in the 1920s in his paper "On the mathematical foundations of theoretical statistics"[21]. Although likelihood and probability are often used as synonyms in daily conversation, both concepts have different meanings in statistics. Probability considers the parameter to be fixed and the data variable, while likelihood considers the data fixed and the parameters variable [22]. In other words, probability is a function of the data given the parameters while likelihood is a function of the parameters given the data and indicates how a particular value for a given parameter is supported by the data.

Assume for example that we want to study the efficiency of a treatment administered to 10 patients. Let X_i represent the outcome for the i^{th} patient with X_i=1 for a success and X_i=0 for a failure and θ represent the proportion of success. Then, the probability density function of the number of successes $S = \sum_{i=1}^{10} X_i$ is a binomial distribution and it can be expressed as $f(s|\theta) = \Pr(S = s|\theta) = \binom{10}{s}\theta^s(1-\theta)^{10-s}$. If we know the proportion of successes of the treatment, θ, and we want to know how many successes there will be in our experiment, then $f(s|\theta)$ is the probability of having s successes in our experiment, given that the proportion of successes is θ. If we know that we have obtained s successes in our experiment and we want to estimate the proportion of successes of the treatment, then $f(s|\theta)$ becomes the likelihood $L(\theta|s)$ that the proportion of successes of the treatment equals θ, given that we have s successes out of 10 in our experiment.

Figure II.1. represents the likelihood function when we obtain S=4 successes in our experiment: $L(\theta|4) = \binom{10}{4}\theta^4(1-\theta)^6$. The value of θ that maximises the likelihood function is called the Maximum Likelihood Estimator (MLE) of θ.

Figure II.1. Likelihood function for the proportion of successes θ, given that we obtain 4 successes in our experiment



b) The Bayes' theorem

Bayes' theorem was first stated by the Reverend Thomas Bayes in the eighteenth century to solve the problem of inferring from experimental data to the parameter, called the inverse probability problem[23-25].

Let A and B be two events. Then, a discrete version of Bayes' theorem for two simple events is:

$$P(B|A) = \frac{P(A|B)\cdot P(B)}{P(A)} = \frac{P(A|B)\cdot P(B)}{P(A|B)\cdot P(B) + P(A|\overline{B})\cdot P(\overline{B})}, \quad (1)$$

where $\overline{B}$ is the complementary event of B so that $P(\overline{B}) = 1 - P(B)$.

Bayesian inference uses Bayes' theorem to combine the information contained in the data with prior knowledge about a given parameter θ to obtain an updated probability for the parameter θ. Bayes' theorem can then be rewritten as follows:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \quad (2)$$

where $p(Y = y|\theta) = p(y|\theta) = L(\theta|y)$ represents the likelihood function for θ, $p(\theta)$ represents the prior probability distribution of the parameter θ and $p(\theta|y)$ represents the posterior distribution for θ.

Equation (2) shows that the posterior distribution $p(\theta|y)$ is proportional to the product of the prior distribution $p(\theta)$ and the likelihood $L(\theta|y)$, as $p(y)$ only depends on the data and is then assumed to be constant. We then have:

$$p(\theta|y) \propto L(\theta|y)p(\theta) \quad (3)$$

c) Choosing the prior distribution

A group of prior distributions, which allows an easy calculation of the posterior distribution, is the conjugate family. A prior distribution $p(\theta)$ belongs to the conjugate family of a likelihood function $L(\theta|y)$, when $p(\theta)$ has the same functional form as $L(\theta|y)$ so that, after applying Bayes' theorem, the posterior distribution $p(\theta|y)$ has the same functional form as $p(\theta)$ [24, 25]. For example, the beta distribution is the conjugate family for the binomial likelihood function, i.e. if we combine a beta prior distribution with a binomial likelihood function, the posterior distribution obtained after applying Bayes' theorem will also be a beta distribution. The specification of the prior distribution parameters depends on the quantity and quality of prior knowledge available. When prior beliefs cannot be represented by conjugate prior distributions, it is

possible to build a discrete prior distribution that matches the belief weights for different values and to interpolate between these weights to get a continuous prior distribution.

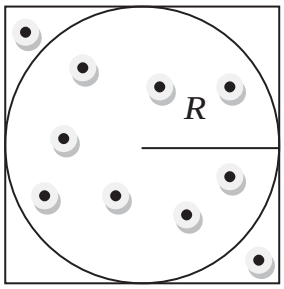
d) Posterior summary measures and posterior sampling

An important advantage of using Bayesian methods is that it is then reasonably straightforward to make inferences on the parameter θ from the posterior distribution, as the probability $P(a < \theta < b|y)$ is easy to compute. Several measures can also be estimated from the posterior distribution to characterize the location and the variability of θ. Among them, the posterior mean $\bar{\theta} = \int \theta\, p(\theta|y) d\theta$ is usually considered as the Bayesian estimate of θ while the posterior variance $\bar{\sigma}^2 = \int (\theta - \bar{\theta})^2\, p(\theta|y) d\theta$ and in particular its square root are used to evaluate the posterior uncertainty about θ. It is also common to define an interval of values [a,b] that are *a posteriori* most credible with probability $(1-\alpha)$, i.e. that verify $P(a \le \theta \le b|y) = 1-\alpha$. This interval is called a $100(1-\alpha)\%$ credible interval.

We can note here that it is necessary to calculate respectively one, two and three integrals to compute the posterior distribution, the posterior mean and the posterior standard deviation. When the prior distribution belongs to the conjugate family, the posterior distribution and posterior summary measures can be obtained through simple calculations. However, in other cases, the integration often becomes complex and it is necessary to determine numerical approximations for the denominator of the posterior distribution and for the posterior summary measures.

A common approach to approximate an integral is called Monte Carlo integration. This method can be illustrated in the example of a circle of radius R circumscribed by a square (Figure II.2)[26]. The area of the circle is πR^2, the area of the square is $4R^2$ and the ratio of the two areas equals $\frac{\pi}{4}$. If we choose random points in the square, then Monte Carlo integration says that the proportion of points which will fall inside the circle will be close to $\frac{\pi}{4}$.

Figure II.2. Illustration of Monte Carlo Integration[26]



The Monte Carlo approach is based on the Strong Law of Large Numbers and the key idea is to approximate the expected value by the sample mean of simulated random variables. In other words, $\mathrm{E}(h(x)) = \int h(x)p(x)\mathrm{dx}$ can be approximated by the finite sum $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$, where $p(x)$ is a probability density function. It can be deduced from this that the posterior mean or Bayesian estimate $\mathrm{E}(\theta|x) = \int \theta p(\theta|x)\mathrm{dx}$ can be approximated by $\frac{1}{n}\sum_{i=1}^{n} \theta_i$, where θ_i are randomly drawn from the posterior distribution. However, when θ is multidimensional ($\theta = (\theta_1, \theta_2, ..., \theta_d)^T$), Monte Carlo integration cannot be used to obtain the marginal posterior distributions and other methods are required.

The Markov Chain Monte Carlo method (MCMC) has caused the expansion of Bayesian inference by enabling the application of Bayesian methods to any multivariate problem[23, 25]. A sequence of vectors $\theta^1, \theta^2, \theta^3, ...$ is a Markov Chain if for any set A, $P\left(\theta^k \in A \middle| \theta^1, ..., \theta^{k-1}\right) = P\left(\theta^k \in A \middle| \theta^{k-1}\right)$. This property, called the Markov property, means that each new observation depends only on the previous observation. The two main MCMC procedures are the Gibbs sampler and the Metropolis-Hastings algorithm.

The Gibbs sampler is based on the property that a multivariate distribution is uniquely determined by its conditional distributions. The algorithm starts by specifying initial values $\theta^0 = \left(\theta_1^0, \theta_2^0, ..., \theta_d^0\right)^T$ for the vector of parameters $\theta = \left(\theta_1, \theta_2, ..., \theta_d\right)^T$. The posterior distribution is then explored by generating $\theta_1^k, \theta_2^k, ..., \theta_d^k \quad (k = 1,2,3,...)$ in a sequential way. At iteration $(k+1)$, the following algorithm is applied:

1) Sample $\theta_1^{(k+1)}$ from $p\left(\theta_1 \middle| \theta_2^k, ..., \theta_{(d-1)}^k, \theta_d^k, y\right)$

2) Sample $\theta_2^{(k+1)}$ from $p\left(\theta_2 \middle| \theta_1^{(k+1)}, \theta_3^k ..., \theta_d^k, y\right)$

…

d) Sample $\theta_d^{(k+1)}$ from $p\left(\theta_d \middle| \theta_1^{(k+1)}, ..., \theta_{(d-1)}^{(k+1)}, y\right)$

The Gibbs sampler therefore generates a sequence of values $\theta_1^1, \theta_2^1, ..., \theta_d^1, \theta_1^2, \theta_2^2, ..., \theta_d^2, ..., \theta_1^3, \theta_2^3, ..., \theta_d^3, ...$ The vectors $\theta^k = \left(\theta_1^k, \theta_2^k, ..., \theta_d^k\right)^T$ create a chain which has the Markov property, i.e. $p\left(\theta^{(k+1)} \middle| \theta^k, \theta^{(k-1)}, ..., y\right) = p\left(\theta^{(k+1)} \middle| \theta^k, y\right)$. The Gibbs sampler is called a Markov Chain Monte Carlo method as it respects the Markov property and uses the Monte Carlo approach to generate sampled values.

Unlike the Gibbs sampler, the Metropolis-Hastings (MH) algorithm directly uses the joint posterior distribution without the need to compute the conditional posterior distributions. Again, the algorithm starts by specifying initial values $\theta^0 = \left(\theta_1^0, \theta_2^0, ..., \theta_d^0\right)^T$ for the vector of parameters $\theta = \left(\theta_1, \theta_2, ..., \theta_d\right)^T$. Then, at iteration $\left(k+1\right)$, the following algorithm is applied:

1) Sample a candidate $\tilde{\theta}$ from a proposal density $q\left(\tilde{\theta}|\theta\right)$, with $\theta = \theta^k$

2) The next value $\theta^{(k+1)}$ will be equal to

- $\tilde{\theta}$ with probability $\alpha\left(\theta^k, \tilde{\theta}\right)$ (accept proposal),
- θ^k otherwise (reject proposal).

with $\alpha\left(\theta^k, \tilde{\theta}\right) = \min\left(r = \frac{p\left(\tilde{\theta}|y\right)q\left(\theta^k|\tilde{\theta}\right)}{p\left(\theta^k|y\right)q\left(\tilde{\theta}|\theta^k\right)}, 1\right)$

The function $\alpha\left(\theta^k, \tilde{\theta}\right)$ is called the "probability of a move".

It has been shown[23, 25] that the Markov Chains produced by the Gibbs sampler and the MH algorithm eventually provide a sample from the posterior distribution and that the obtained summary measures consistently estimate the true posterior summary measures. An initial portion of the Markov Chain samples called "burn-in period" is usually discarded in order to reduce the effects of initial values on the posterior inference.

The application of Bayesian analyses has grown in epidemiological studies in the last decades. Examples of these analyses include Bayesian approaches to model prior and posterior distributions for odds ratios in case-control studies[27-33] and to analyse survival data [34-36]. In this thesis I have focussed on Bayesian shrinkage methods to

address the random error in the estimation of multiple associations and on Bayesian methods to address misclassification bias by considering it as a missing data problem. The methods I chose to apply are only a small selection of the Bayesian methods available in the literature. However, the objective here was to use methods that can be implemented with standard statistical software and to make them easily accessible by epidemiologists.

In the following two subsections I will give an overview on Bayesian shrinkage methods and Bayesian approaches to address bias by considering it as a missing data problem.

2. Shrinkage methods

a) The shrinkage principle

Let us consider the estimation of the average risk θ of an outcome in a single cohort[37]. If we denote N the size of a sample and A the number of cases in this sample, the observed proportion A/N is the usual Maximum Likelihood Estimator (MLE) of the risk parameter θ under the assumption that the observed proportion A/N differs from the target risk θ in only a random fashion, in accord with the assumed probability model. Under the previous assumption, we have the following equation:

$$E(A/N) = \theta \tag{4}$$

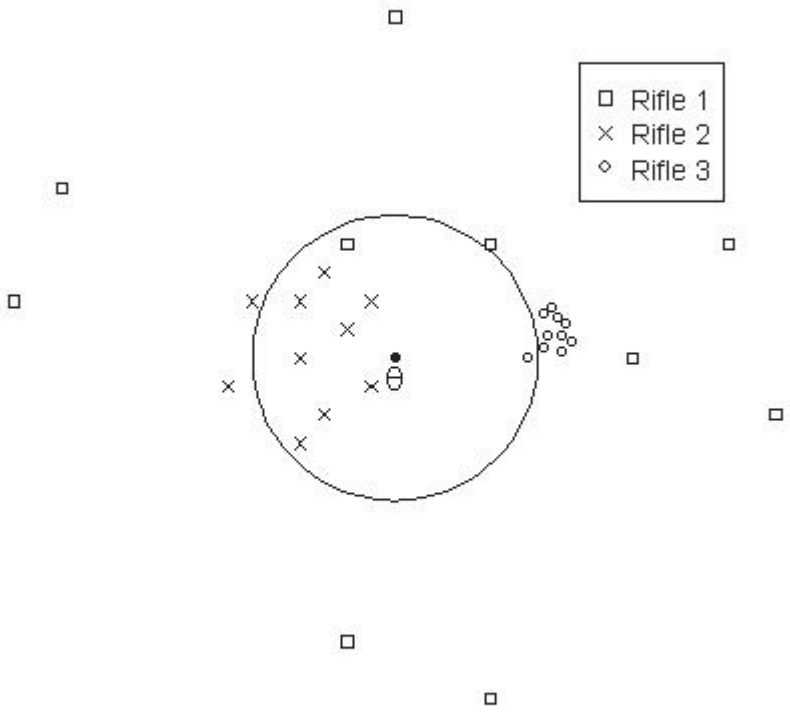
It means that the proportion A/N is an unbiased estimate for the average risk θ. Unbiasedness is a property of a good estimator. However, it is not the only one. We have to take into account the scatter of the estimator too. Estimators with large standard deviations are unreliable estimators of the target parameter, even if they are unbiased.

The problem is that if we want to reduce the standard deviation of an estimator, we may increase its bias.

The following example illustrates the principle that bias and scatter have to be considered together.

Suppose we have three rifles and we want to test which one will more often ‘win’ when sighted on θ and shot, where a ‘win’ is a hit anywhere inside the ring around θ.

Figure II.3. The rifle example (1) - Illustration of bias and scatter[37]



Rifle 1 is like an estimator with no bias but large random error; Rifle 2 is like an estimator with moderate bias and moderate random error; and Rifle 3 is like an estimator with large bias but small random error. We can see here that Rifle 2 wins because it has the smallest average distance from target θ. This explains why a common criterion of accuracy of an estimator (which takes into account both bias and scatter) is the average squared distance from θ called Expected Squared Error (ESE). One very useful property of ESE is: $ESE = bias^2 + SD^2$, where SD is the standard deviation of the estimator, i.e. the scatter of the estimator around its expectation across random samples from the population.

This equation displays the trade-off we must make between bias and scatter when choosing estimators based on their expected squared error.

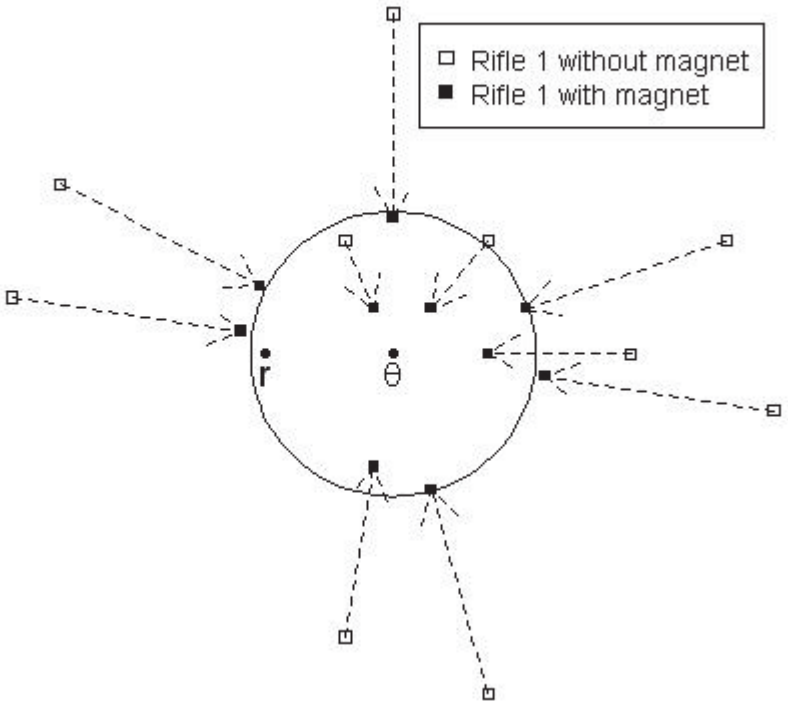
Now, let us turn back to our estimator A/N. It is unbiased and has the smallest standard deviation possible for such an estimator. Consequently, if we want to find an estimator with a smaller ESE, we have to search among biased estimators with a smaller standard deviation than A/N. One systematic approach to perform this search is the Bayes estimation. Before seeing the study data, let us suppose that our best guess is that the average risk θ is somewhere near a given value r and, as we are unsure about our guess, let us suppose that we would give our best guess the same weight as we would give to the results from a study of size n. The values r and n are prior parameters. In particular, r is called the prior mean for θ, and n is called the prior effective sample size.

A reasonable approximate Bayes estimator can often be obtained by taking a weighted average of the ordinary estimator and our prior guess for θ. One such estimator weighs

the observed proportion A/N and the prior mean r by their respective sample sizes N and n. If we let $w = N/(N+n)$, this estimator is the weighted average : $\hat{b} = w(A/N) + (1-w)r$, and is considered as a posterior mean for θ.

Why would we expect $\hat{b}$ to do better than A/N when n is greater than 0? If we refer to our unbiased but very scattered Rifle 1, which corresponds to A/N, let us imagine that Rifle 1 shoots steel bullets and that we place a powerful magnet at r as the next figure shows.

Figure II.4. The rifle example (2) - Illustration of shrinkage[37]



With this magnet, the hits are now biased away from θ, but less scattered and closer to θ on average (since more hits fall into the ring), provided r is not too far from θ. The same phenomenon applies with the estimator $\hat{b}$. As we increase n, we decrease w and so increase the weight we put on the prior guess r, which increases the bias but reduces the scatter of $\hat{b}$. Of course, the better our guess, the closer r to θ, and the less unbiased $\hat{b}$.

The estimator $\hat{b}$ is an example of a 'shrinkage estimator' because it is the usual estimator A/N 'shrunk' toward the point r by a degree proportional to $1-w$. Empirical-Bayes and semi-Bayes methods, ridge regression, penalized estimation, Stein estimation, and hierarchical regression are all different types of shrinkage methods[37, 38].

b) Application of shrinkage in spatial epidemiology

Disease mapping is one of the earliest applications of Bayesian shrinkage in Epidemiology. Environmental studies often explore the patterns of disease in order to assess which potential areas are at low/high risk of disease. When counts of cases are available for each studied geographical area, the usual model assumes that these observed counts $\{y_i, i = 1,...,n\}$ are independent Poisson random variables with parameters $\{\lambda_i = e_i\theta_i, i = 1,...,n\}$, where e_i are the expected counts for the i^{th} area and θ_i are the relative risks for the i^{th} area[23, 39-41]. However, for small areas, the estimates for θ_i can be very unstable due to small numbers. Shrinkage methods can help reducing random noise by smoothing these unstable estimates across neighbouring

areas. A standard shrinkage model is the Besag, York and Mollié (BYM) convolution model[42], which takes the following form:

$$\log(\theta_i) = x_i^T \beta + v_i + u_i ,$$

where x_i^T is the ith row of a covariate design matrix, β is a vector of regression parameters, v_i represents the uncorrelated heterogeneity that allows for overdispersion and u_i represents the spatially correlated heterogeneity. In a Bayesian analysis, the components of the model are then assigned prior distributions and posterior sampling of these parameters can be performed via MCMC algorithms.

c) Semi-Bayes and empirical Bayes adjustments towards the global mean (applied in Chapters III, IV and V)

In this subsection I describe the principles of Semi-Bayes and Empirical Bayes adjustments towards the global mean to address the issue of multiple comparisons and I illustrate the effects of these two methods in an occupational study of lung cancer. This subsection represents a development of work which was completed before the formal commencement of my PhD[17]. It represents a very preliminary exploration of the issues and practicalities of conducting such analyses, and I therefore include it as 'background' rather than as a separate chapter.

In epidemiological studies, Empirical Bayes (EB) and Semi-Bayes (SB) adjustment methods have been shown to be more valid approaches to the problem of multiple comparisons than Bonferroni method, particularly when the parameters to be estimated can be divided into groups within which they are similar or "exchangeable" on the basis of *a priori* knowledge[43]. Thus, Empirical Bayes and Semi-Bayes methods can enable the avoidance of numerous false positive associations, and can produce effect estimates that have smaller expected squared errors.

i. Material and methods

The aim of Bayesian estimation is to reduce the expected squared error of an estimator[37]. The method consists in giving a prior expectation to the estimated parameter. Then a posterior estimate is calculated as a weighted mean of the standard estimate and the prior expectation. Consequently, if the prior expectation is not too far from the true parameter, the expected squared error and the probability of type I error are reduced. EB and SB adjustment methods assume that the observed variation of the estimated parameters (e.g. odds ratios) around their global mean is larger than the variation of the true parameters[43]. The EB method aims at estimating the variation of the true parameters directly from the data, whereas the SB method specifies an *a priori* value for the variation of the true parameters so that they have a reasonable range of variation (e.g. a Var_{true} of 0.25 implies that 95% of the true relative risks are within a 7-fold range of each other[43]). The «adjustment» then consists of shrinking outlying estimates towards the geometric mean of the estimates' distribution. The larger the individual variance of the estimates, the stronger is the shrinkage, so that the shrinkage is stronger for less reliable estimates based on small numbers. The effect of shrinkage is to reduce the overall variance of the estimates. The EB method estimates the variance of the true log odds ratios (Var_{true}) as:

$$Var_{true} = Var_{obs} - Var_{mean} \quad [1]$$

where Var_{obs} is the observed sample variance of the log odds ratios estimates, and Var_{mean} is the mean of the estimated variances of each log odds ratio estimate. Since Var_{true} must be a positive value, Var_{obs} must be greater than Var_{mean}. If the estimated variances do not satisfy this inequality, the SB method, in which Var_{true} is set by the investigator, should be used instead of the EB method. EB and SB adjustments can be validly used under specific conditions. Firstly, the distribution of the estimates to be

adjusted must be well approximated by a log normal distribution. Secondly, if the exposures are quantitative, they must be rescaled so that the log odds ratios for a one-unit increment of exposure must be comparable. Finally, if there are prior associations between the odds ratios, these must be taken into account[43]. We have considered the simplest case in which there are no such associations or they can be neglected. We applied the EB and SB adjustment methods to a case-control study of occupational risk factors for lung cancer[44].This study was carried out between 1990 and 1992 in two areas of Italy: Turin and the Eastern part of Veneto region. Cases (956 men and 176 women) were all individuals with incident primary lung cancer, aged less than 75 and resident in the study areas. Controls (1,253 men and 300 women) were randomly selected from the local population registries and frequency matched with cases by gender, study area and five-year age groups. Information on basic demographic details, active and passive smoking, and a lifetime occupational history was collected. In particular, the dates of beginning and ending work, as well as the job title and branch of industry, were recorded for each occupational period that lasted at least 6 months. Job titles and branches of industry were coded blind to case-control status according to the International Standard Classification of Occupations (ISCO)[45] and the International Standard Industrial Classification (ISIC)[46], respectively. These classifications, based on a maximum number of 5 and 4 digits, respectively, increase the specificity of each occupation/industry with increasing number of digits (e.g., ISCO code 93: painters, and ISCO code 93190: structural steel and ship painters). A logistic regression model was built for each job and for each industry, for men and women separately. Covariates included in the models were age, study area, and cigarette smoking status. Odds ratios of lung cancer with corresponding 95% confidence intervals (CI) were estimated for all jobs and industries and SB and EB adjustments were applied to the obtained estimates.

The computation of EB estimates was often impossible because the estimate of Var_{true} was negative (see equation 1 above). Even when EB adjustment was possible, the estimated Var_{true} was very small, resulting in a very strong shrinkage towards 1 for all ORs and the "weeding out" of potentially true associations[43].We therefore used SB adjustment in the analyses. In this paper, we show the effects of SB adjustment on the estimates of ORs of lung cancer for job titles among men. Analyses were conducted using SAS and R software. The codes for the SB adjustments are available in Appendix III.

ii. Results

Figure II.5 shows the lower bound of the confidence intervals of the SB-adjusted ORs of lung cancer against the lower bound of the confidence intervals of the standard (not SB-adjusted) ORs for job titles associated with an OR above 1. The figure thus depicts the changes in the statistical significance of risk estimates produced by SB adjustment. In the lower left quadrant are increased risk estimates that did not reach statistical significance with either the standard unadjusted estimation nor after SB adjustment (probable true negatives). In the lower right quadrant are risk estimates that did reach statistical significance in the standard unadjusted estimation, but lost it after SB adjustment (probable false positives). The upper right quadrant contains risk estimates that were statistically significant before and after SB adjustment (probable true positives). As expected, the upper left quadrant is empty, i.e. there were no increased risk estimates that gained statistical significance through SB adjustment (probable false negatives). Both the probable elimination of false positives and the shrinkage of estimates produced by the SB adjustment can be observed in Figure II.5. The probable elimination of false positives is shown by the fact that only some of the statistically

significant findings remained so after SB adjustment. Shrinkage is shown by the fact that SB-adjusted estimates depart from a 45-degree line to be pulled towards the centre of the distribution.

Figure II.5. Scatter plot of the lower bound of the Semi-Bayes (SB) adjusted 95% confidence intervals (CI) against the lower bound of the standard 95% CI for increased odds ratios (OR) of lung cancer for different job titles, defined on the basis of 2, 3, 4 and 5 ISCO digits[45]. Men

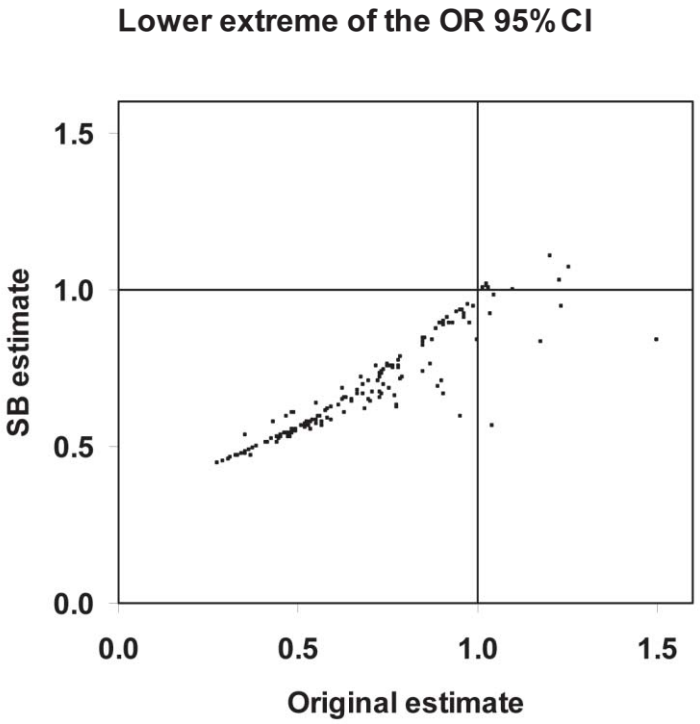


Table II.2 shows the numbers of statistically significant estimates expected by chance, observed without adjustment for multiple comparisons and observed after Bonferroni correction and SB adjustment. Bonferroni adjustment is strongly penalising since it removes all statistically significant estimates. SB adjustment, on the other hand, can be used to identify the most robust findings for further investigation. For 3-digit ISCO

codes, for example, if a decision rule was based on an elevated OR with p <0.05, in our study we would further investigate 6 occupations among men using unadjusted results, versus 2 using SB adjustment.

Table II.2. Frequencies of statistically significant increased risks of lung cancer for job titles (defined on the basis of 1 to 5 digit ISCO codes[45]) before and after Bonferroni and Semi-Bayes adjustments. Men

Number of digits of ISCO codes	Number of OR[a] estimates	Number of statistically significantly increased ORs[a]			
		Expected*	Observed		
			No adjustment	Bonferroni adjustment	Semi-Bayes adjustment
1	10	0.25	1	0	1
2	61	1.53	4	0	4
3	122	3.05	6	0	2
4-5	154	3.85	5	0	2

[a] OR, odds ratio
* based on α =0.05

iii. Discussion

SB and EB shrinkage methods are valid and robust methods for addressing the problem of multiple comparisons in occupational studies. Our findings show that the SB method, by reducing the variability of the estimates, may decrease the number of false positive findings, while not eliminating all positive findings. A potentially interesting application of SB method in occupational epidemiology is the analysis of the numerous situations of exposure identified by the combination of ISCO and ISIC codes. Rothman[47] pointed out that deciding to not making Bonferroni adjustments for multiple comparisons may be preferable because it leads to fewer errors of interpretation, given that the observed data are real observations on a natural phenomenon and not random numbers. Moreover, when dealing with inferential investigations, statistical significance is just one of the several components of causal

inference. However, Bayesian adjustments for multiple comparisons avoid the problems associated with the Bonferroni method, while producing effect estimates that are, on the average, more valid than the unadjusted estimates. Furthermore, they can be easily carried out using standard statistical software. R codes are available in Appendix III. Thus, their use is recommended in exploratory analyses.

iv. Calculation of the SB/EB estimates

Let us consider a group of n exchangeable odds ratios (OR) estimates, an "ensemble" of ORs, each with an estimated variance, and let us denote *i* individual members of this ensemble. Taking logarithms, we derive a weighted average, as shown below[43, 48].

$$\ln(\mathrm{OR})_{mean} = \frac{\sum(\ln(\mathrm{OR})_i * w_i)}{\sum w_i}$$

The weights have the form:

$$w_i = \frac{1}{S_i^{\,2} + Var_{true}},$$

where $S_i^{\,2}$ is the variance of each $\ln(\mathrm{OR})_i$.

In a SB analysis, Var_{true} is specified arbitrarily at the start.

In an EB analysis, Var_{true} is estimated from the data. Thus EB analyses require the use of iteration, where an initial guess of Var_{true} is used, and then this initial guess is refined iteratively. In more detail, let:

$$D = \ln(\mathrm{OR})_i - \ln(\mathrm{OR})_{mean};$$

$$\hat{V}_{obs} = \frac{\sum(w_i * D_i^{\,2})}{\sum w_i}, \text{ where } \hat{V}_{obs} \text{ is our estimate of } Var_{obs};$$

$$\hat{V}_{mean} = \frac{\sum(w_i * S_i^{\,2})}{\sum w_i}, \text{ where } \hat{V}_{mean} \text{ is our estimate for } Var_{mean}.$$

We can now estimate Var_{true} by $\hat{V}_{true} = \hat{V}_{obs} - \hat{V}_{mean}$ or by $\max(\hat{V}_{obs} - \hat{V}_{mean}, \tau^2)$, where τ^2 is a user-specified minimum plausible value for Var_{true}.

Finally, we can derive the SB/EB estimate of each ln(OR) as a weighted average of the original estimate and the mean of the estimates as described below.

$$EB_i = \frac{\hat{V}_{true} * \ln(\mathrm{OR})_i + S_i^{\,2} * \ln(\mathrm{OR})_{mean}}{\hat{V}_{true} + S_i^{\,2}}$$

$$SB_i = \frac{Var_{true} * \ln(\mathrm{OR})_i + S_i^{\,2} * \ln(\mathrm{OR})_{mean}}{Var_{true} + S_i^{\,2}}$$

We can note that if $\hat{V}_{true}$ (for EB estimates) or Var_{true} (for SB estimates) is large, this gives more weight to the original estimate. On the other hand, if the variance of the individual estimate $S_i^{\,2}$ is large, this gives more weight to the overall mean of the estimates.

Then, if we denote $E = w_i * S_i^{\,2} * D_i * \sqrt{\frac{V_{mean}}{S_i^{\,2}}}$, the variance of each adjusted estimate is

$$S_i^{\,2} * \left(1 - S_i^{\,2} * w_i\right) + \frac{2 * E^t E}{n}.$$

d) Bayesian hierarchical regression (applied in Chapter V)

Bayesian hierarchical regression (HR) is a particular type of multilevel modelling, a technique which specifies several levels of associations among the variables in a study and which has given birth to several methods in Epidemiology such as random effect models or mixed models[37].

Let us consider a standard logistic regression model estimating a group of *n* ORs[49-57].

$$\text{logit}\left[P\left(Y=1|\mathbf{X},\mathbf{W}\right)\right]=\alpha+\mathbf{X}\beta+\mathbf{W}\gamma \quad (5)$$

where Y is the disease status, $\mathbf{X}$ is the matrix of exposures of interest (each row of $\mathbf{X}$ gives information on the exposures of interest for one individual), $\mathbf{W}$ is a matrix of potential confounders, α is the intercept term and β and γ are the column vectors of logistic regression coefficients. While the SB and EB adjustments described in the last subsection shrink the estimates towards their global mean, HR shrinks the OR estimates towards prior means which are obtained through a second-stage model, as follows[52]:

$$\beta=\mathbf{Z}\pi+U\,. \quad (6)$$

$\mathbf{Z}$ is a $n\times p$ matrix indicating the relationship between the n exposures of interest and p second-stage covariates.

π is a p-element vector of the coefficients corresponding to the effects on the disease of the second-stage covariates.

U is a n-element vector of error terms with $U \sim N\left(0,\tau^2\mathbf{T}\right)$.

Three approaches can be used to compute the prior means $\mathbf{Z}\pi$ of the ln(OR)s:

- The fully Bayesian approach requires to specify distribution for both π and τ^2.
- The parametric empirical Bayes approach estimates both π and τ^2 from the data using iterations.
- The semi-Bayes approach requires specifying a value for τ^2 but estimates π.

In the semi-Bayes approach, which is used in Chapter V, the second-stage coefficients π are estimated through weighted least squares with:

$$\tilde{\pi} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\hat{\beta}$$

where $\mathbf{W} = \left[\hat{\mathbf{V}} + \tau^2\mathbf{T}\right]^{-1}$ and $\hat{\mathbf{V}}$ is a diagonal matrix composed of the estimated variances of the first-stage coefficients estimated in (5) $\hat{\beta}$.

HR estimates $\hat{\beta}_{HR}$ are then obtained by averaging the ln(OR) estimates $\hat{\beta}$ with their respective prior means $\mathbf{Z}\tilde{\pi}$:

$$\hat{\beta}_{HR} = \mathbf{B}\mathbf{Z}\tilde{\pi} + (\mathbf{I} - \mathbf{B})\hat{\beta}, \text{ where } \mathbf{B} = \mathbf{W}\hat{\mathbf{V}} = \left(\hat{\mathbf{V}} + \tau^2\mathbf{T}\right)^{-1}\hat{\mathbf{V}}$$

Their covariance matrix is estimated by $\tilde{\mathbf{C}} = \hat{\mathbf{V}}\left(\mathbf{I} - (\mathbf{I} - \mathbf{H})'\mathbf{B}\right)$ where:

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}$$

In the particular case where $p = 1$ (i.e. the matrix $\mathbf{Z}$ is constituted of only one column) and all the elements of $\mathbf{Z}$ are equal to 1, HR shrinks the ln(OR)s towards the global mean and becomes then identical to the SB and EB adjustments towards the global mean described previously.

Models (5) and (6) can also be combined using a mixed-effects logistic model:

$$\text{logit}\left[P\left(Y=1|\mathbf{X},\mathbf{W}\right)\right]=\alpha+\mathbf{XZ}\pi+\mathbf{X}U+\mathbf{W}\gamma$$

where π and γ are column vectors of fixed coefficients and U is a column vector of a random coefficient[55, 56]. However, in Chapter V, the *n* ORs are estimated separately by *n* first-stage logistic regression models and hierarchical regression therefore needs to be applied in two steps.

In the last subsection and in Chapters III, IV and V, I refer to the Semi-Bayes intervals as "confidence intervals" rather than "credibility intervals" in order to be consistent with the literature[43, 51, 58-61] and because these intervals are not obtained directly from posterior distributions but computed from the posterior means and variances. For the same reasons and in order to allow the comparison with standard Maximum Likelihood estimates, I also define a SB OR estimate as "statistically significant" when its interval does not include 1.

3. Bayesian methods for the analysis of bias

All of the Bayesian methods illustrated so far have attempted to address the problem of random error, particularly in studies involving multiple comparisons. In this section, we consider Bayesian methods for addressing systematic error or bias.

a) Bias analysis via imputation of missing data

All forms of bias can be seen as involving missing data[62]. To illustrate this, let us consider the following sample:

id	*Y*	*T*	*X*	*Z*	*S*
1	1	0	0	1	1
2	0	0	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	0	0	0	1	1
6	1	0	0	0	1
7	0	1	1	1	1
8	0	1	0	0	1
9	0	0	1	1	1
10	1	1	0	1	1

where *Y* is a disease outcome, *T* is the exposure of interest, *X* is the exposure measurement, *Z* is a potential confounder and *S* is the indicator for the selection of the subject. For example: *Y*: lung cancer, *T*: smoking, *X*: reported smoking, *Z*: occupational exposures, *S*: eligibility of the subject to participate in the study.

Then, if internal validation data is available, selection bias, unmeasured confounding and information bias could be represented respectively by scenarios 1, 2 and 3 below, where the information given by internal validation data is given in bold.

Scenario 1

Selection bias: data are partially observed for 4 non-selected subjects (validation data).

id	*Y*	*T*	*X*	*Z*	*S*
1	**1**	.	.	.	**0**
2	**0**	.	.	.	**0**
3	**1**	.	.	.	**0**
4	**1**	.	.	.	**0**
5	0	0	0	1	1
6	1	0	0	0	1
7	0	1	1	1	1
8	0	1	0	0	1
9	0	0	1	1	1
10	1	1	0	1	1

Scenario 2

Unmeasured confounding: the confounder is missing for 7 subjects and observed for 3 subjects (validation data).

id	*Y*	*T*	*X*	*Z*	*S*
1	1	0	0	.	1
2	0	0	1	.	1
3	1	1	0	.	1
4	1	1	1	.	1
5	0	0	0	**1**	1
6	1	0	0	**0**	1
7	0	1	1	.	1
8	0	1	0	**0**	1
9	0	0	1	.	1
10	1	1	0	.	1

Scenario 3

Information bias: Information on misclassified exposure (or a surrogate of exposure) is available for all subjects, but information on the true exposure of interest is missing for most (7 subjects) and only available for a validation subsample (3 subjects).

id	*Y*	*T*	*X*	*Z*	*S*
1	1	.	0	1	1
2	0	.	1	0	1
3	1	.	0	1	1
4	1	.	1	0	1
5	0	**0**	0	1	1
6	1	**0**	0	0	1
7	0	.	1	1	1
8	0	**1**	0	0	1
9	0	.	1	1	1
10	1	.	0	1	1

If validation data is not available, the data in bold is missing. Thus, in scenario 1, data become totally missing for 4 subjects. In scenario 2, the confounder becomes missing for all 10 subjects and in scenario 3, the exposure of interest becomes unknown for all 10 subjects.

In each of these scenarios, missing values can be of three types[63] :

- missing completely at random (MCAR): Missingness is independent of observed and unobserved data. In this case, it is possible to carry out a naïve analysis, in which only the fully observed individuals for scenario 1 and the fully observed variables for scenario 2 and 3 are included, as it will yield no bias.
- missing at random (MAR): Missingness depends only on the observed data, which means that data is MCAR conditional on the observed variables responsible for missingness. Missing data can then be imputed conditioning on this set of variables.

- missing not at random (MNAR): Missingness depends on the unobserved data and the missingness mechanism needs to be included in the analysis.

I will focus here on MAR data. In the presence of internal validation data, the association between missing data and the variables responsible for missingness can be estimated from the validation subsample. Many methods have been developed to handle partially missing data. These methods include multiple imputation, data augmentation, the "Expectation-Maximisation" (EM) algorithm, and inverse probability weighting[63-69]. In Section 3, I focus on the approach of multiple imputation because it is now implemented in most commonly used software in Epidemiology and it is therefore accessible to most researchers in this area.

In the absence of internal validation data, prior assumptions must be made about the association between missing data and the variables responsible for missingness. Corrected estimates can then be obtained by conducting a sensitivity analysis with these prior assumptions.

In the following subsections I will give a brief overview on multiple imputation and the different types of sensitivity analysis.

b) Multiple imputation (applied in Chapter VI)

The principle of multiple imputation (MI) is to use the data from individuals who have complete information on all variables (internal validation data) to estimate the association between partially missing and fully observed variables. Then, this association is used to complete the whole dataset by drawing the missing values from the distribution of the partially missing variables given the fully observed variables.

Let us go back to scenario 3 of our previous example. The model of interest would be:

$$\text{logit}\left(P(Y)\right) = \beta_0 + \beta_T T + \beta_Z Z + \varepsilon, \text{ where } \varepsilon \sim N\left(0, \sigma^2 \text{Id}\right) \text{and Id is the identity matrix.}$$

A possible imputation model which would estimate the association between the partially missing variable T and the other variables in the validation data would be:

$$\text{logit}\left(P(T)\right) = \lambda_0 + \lambda_X X + \lambda_Y Y + \lambda_Z Z + e, \tag{7}$$

where $e \sim N\left(0, \gamma^2 \text{Id}\right)$ and Id is the identity matrix.

The MI algorithm draws the missing values for T from (7) to obtain a completed dataset for K times so that K different completed datasets are created[69, 70]. The model of interest is then fitted to each of the K datasets, and K ln(OR) estimates $\hat{\beta}_T^k, k = 1,..., K$ with respective variances $\hat{V}_k, k = 1,..., K$ are computed. According to Rubin rules[71], the MI estimator can then be calculated as the average of the K ln(OR) estimates: $\hat{\beta}_T^{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_T^k$.

The variance of the MI estimator is computed as a weighted average of the between and within-imputations variability:

$$\hat{V}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{V}_k + \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}\left(\hat{\beta}_T^k - \hat{\beta}_T^{MI}\right)^2$$

The within-imputations variability $W = \frac{1}{K}\sum_{k=1}^{K}\hat{V}_k$ indicates the variability of the ln(OR) estimates if it were obtained from one complete dataset whereas the between-imputations variability $B = \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}\left(\hat{\beta}_T^k - \hat{\beta}_T^{MI}\right)^2$ specifies the variability of the ln(OR) estimates across the imputations.

Two main types of uncertainty need to be taken into account in order to give 'proper' imputations[71, 72]:

1. The uncertainty about the distribution of the missing values, which is reflected by the variance of the error term $\gamma^2 \mathrm{Id}$ in the imputation model. At each imputation, the missing values then need to be drawn from the distribution $P(T|X)$.
2. The uncertainty about the imputation model coefficients, which is illustrated by the variance of the estimates of $\lambda_0, \lambda_X, \lambda_Y, \lambda_Z$ when the imputation model is fitted to the validation data. At each imputation, $\lambda_0, \lambda_X, \lambda_Y, \lambda_Z$ need to be drawn from their posterior predictive distribution.

The second condition can be interpreted from a Bayesian point of view[70, 73]: Let us denote: $\Theta = (\beta_0, \beta_T, \beta_Z, \lambda_{0,} \lambda_X, \lambda_Y, \lambda_Z, \gamma^2)$ and $\mathrm{Data}_{\mathrm{obs}}$ and $\mathrm{Data}_{\mathrm{mis}}$, respectively the observed and missing data. Then the joint posterior distribution of Θ and the missing data given the observed data can be given by:

$P(\mathrm{Data}_{\mathrm{mis}}, \Theta|\mathrm{Data}_{\mathrm{obs}}) = P(\mathrm{Data}_{\mathrm{mis}}|\mathrm{Data}_{\mathrm{obs}})P(\Theta|\mathrm{Data}_{\mathrm{obs}}, \mathrm{Data}_{\mathrm{mis}})$ and the marginal posterior distributions are given by:

$P(\mathrm{Data}_{\mathrm{mis}}|\mathrm{Data}_{\mathrm{obs}}) = \int P(\mathrm{Data}_{\mathrm{mis}}|\mathrm{Data}_{\mathrm{obs}}, \Theta)P(\Theta|\mathrm{Data}_{\mathrm{obs}})d\Theta$ and

$P(\Theta|\mathrm{Data}_{\mathrm{obs}}) = \int P(\Theta|\mathrm{Data}_{\mathrm{obs}}, \mathrm{Data}_{\mathrm{mis}})P(\mathrm{Data}_{\mathrm{mis}}|\mathrm{Data}_{\mathrm{obs}})d\mathrm{Data}_{\mathrm{mis}}$.

In order to perform Bayesian multiple imputations in our example, we then need to respect the following steps:

1- Choose a prior distribution for Θ.

2- Compute the posterior distribution for Θ given the observed data.

3- Sample $\lambda_0, \lambda_X, \lambda_Y, \lambda_Z$ and γ^2 from their respective posterior distributions and impute T.

4- Estimate Θ in the completed dataset.

5- Repeat steps 3 and 4 K times replacing each time the posterior distributions with the estimates obtained in 4 to create the K imputed datasets.

However, other approximate Bayesian imputation algorithms such as imputation-coefficient resampling[74], which draws $\lambda_0, \lambda_X, \lambda_Y, \lambda_Z$ from the distribution estimated in the validation data at each imputation are also considered 'proper' according to Little and Rubin[63].

c) Sensitivity analysis (applied in Chapter VI)

Sensitivity analysis, or quantitative bias analysis, updates the conventional estimate of an association according to several parameters, called bias parameters[20]. In the case of selection bias, these parameters are the response rates or proportions of participating subjects within categories of outcome and exposure. In the case of unmeasured confounding, the bias parameters are the association between the exposure and the unmeasured confounder and the association between the disease and the unmeasured confounder. Finally, in the case of measurement error, the bias parameters define the association between the observed exposure and the true exposure (e.g. the sensitivity and the specificity in the case of a binary exposure).

When internal validation data are available, information on bias parameters can be obtained from those data. When such data are not available, *a priori* values have to be given to these parameters. According to the type of prior information that we gather,

different bias analysis techniques can be applied. Table II.3 summarises the main characteristics of these techniques.

In the simple bias analysis, the estimate obtained in the data is corrected for bias only once and the output is a single corrected estimate, which does not account for random error[20]. Multidimensional bias analysis (also called fixed-parameter bias analysis) applies simple bias analysis several times in order to represent different possible bias scenarios. It provides several values for the corrected estimate but does not give any indication on which values are more plausible and does not address random error either. Probabilistic sensitivity analysis assigns probability distributions to the bias parameters and, by repeatedly sampling from these distributions, it yields a frequency distribution for the corrected estimate which also incorporates random error. Unlike simple and multidimensional bias analysis techniques, probabilistic sensitivity analysis gives then information on central tendency and uncertainty of the corrected estimate. There is a debate on whether probabilistic sensitivity analysis should be considered as Bayesian. It has been argued that probabilistic sensitivity analysis is a semi-Bayes approach in which prior distributions can be used for some parameters but not for others and in which the prior distributions are not correctly updated with the observed data[20, 62, 75]. On the contrary, a fully-Bayesian sensitivity analysis assigns prior distributions to all the parameters (bias parameters and association of interest) and combines these prior distributions with the data to yield posterior distributions for all the parameters. Fully-Bayesian sensitivity analysis is generally conducted by using an MCMC algorithm[76, 77]. However, an alternative approach to this algorithm is to translate the prior distributions into new data records and to then apply methods to correct for partially missing data[62].

Table II.3. Characteristics of several quantitative bias analysis techniques[20]

Bias analysis technique	Prior information	Output for the corrected estimate of the association of interest
Simple bias analysis	One single prior value for each bias parameter	One single corrected estimate
Multidimensional bias analysis	Range of prior values for each bias parameter	Range of corrected estimates
Probabilistic sensitivity analysis	Prior probability distributions for all bias parameters (and in some cases for the estimate of the association of interest)	Probability distribution of the corrected estimate
Fully Bayesian sensitivity analysis	Prior probability distributions for all bias parameters and for the estimate of the association of interest	Probability distribution of the corrected estimate

SECTION 2

Random error

CHAPTER III

Lung cancer and occupation: A New Zealand cancer registry-based case-control study

Marine Corbin, David McLean, Andrea 't Mannetje, Evan Dryson, Chris Walls, Fiona McKenzie, Milena Maule, Soo Cheng, Chris Cunningham, Hans Kromhout, Aaron Blair and Neil Pearce

Background There are many proven and suspected occupational causes of lung cancer, which will become relatively more important over time, as smoking prevalence decreases.

Methods We interviewed 457 cases aged 20–75 years notified to the New Zealand Cancer Registry during 2007–2008, and 792 population controls. We collected information on demographic details, potential confounders, and employment history. Associations were estimated using logistic regression adjusted for gender, age, ethnicity, smoking, and socio-economic status.

Results Among occupations of a priori interest, elevated odds ratios (ORs) were observed for sawmill, wood panel and related wood-processing plant operators (OR 4.63; 95% CI 1.05-20.29), butchers (OR 8.77, 95% CI 1.06-72.55), rubber and plastics products machine operators (4.27; 1.16-15.66), heavy truck drivers (2.24; 1.19-4.21) and workers in petroleum, coal, chemical and associated product manufacturing (1.80; 1.11–2.90); non-significantly elevated risks were also observed for loggers (4.67; 0.81-27.03), welders and flame-cutters (2.50; 0.86-7.25), pressers (5.74; 0.96-34.42), and electric and electronic equipment assemblers (3.61; 0.96-13.57). Several occupations and industries not of a priori interest also showed increased risks, including nursing associate professionals (5.45; 2.29-12.99), enrolled nurses (7.95; 3.10-20.42), care givers (3.47; 1.40-8.59), plant and machine operators and assemblers (1.61; 1.20-2.16), stationary machine operators and assemblers (1.67; 1.22-2.28), food and related products processing machine operators (1.98; 1.23-3.19), labourers and related elementary service workers (1.45;

1.05-2.00), manufacturing (1.34; 1.02-1.77), car retailing (3.08; 1.36-6.94), and road freight transport (3.02; 1.45-6.27).

Conclusions Certain occupations and industries have increased lung cancer risks in New Zealand, including wood workers, metal workers, meat workers, textile workers and drivers.

Key words: lung cancer; occupation; case-control study; wood workers; metal workers; welding; machine operators; food process workers

Chapter based on American Journal of Industrial Medicine 2011; 54(2):89–101

Introduction

Lung cancer is the leading cause of cancer mortality in the world[78]. In New Zealand, it is currently the most frequent cause of cancer death for men and it is projected to be the same for women by 2011[79]. Tobacco smoke has been established as the major risk factor for lung cancer since the 1950s, and the increases in tobacco consumption since the First World War have resulted in major increases in lung cancer incidence[80]. Although tobacco is the major risk factor for lung cancer globally, many other exposures can also increase lung cancer risk, both in smokers and non-smokers. In particular, a number of occupational exposures are known to be associated with this tumour, including asbestos, radon, inorganic arsenic, chromium, nickel, and polycyclic aromatic hydrocarbons[81]. Many professional activities still entail exposure to these substances, and other occupational exposures are suspected lung carcinogens. For example, in New Zealand, an increased risk of lung cancer has been found in several different occupations including meat workers[82] and wood workers[83].

We have conducted a population-based case-control study of lung cancer in New Zealand in order to assess whether previously reported occupational associations persist and to identify other occupations and occupational exposures that may also contribute to

lung cancer risk. This study is part of an ongoing series of cancer registry-based case-control studies investigating occupational cancer in New Zealand which has also included studies of non-Hodgkin's lymphoma (NHL), bladder cancer and leukaemia[59-61]. We report here the lung cancer findings by occupation and industry.

Methods

The general methodology has been described in reports on previous case-control studies in the series[59-61], and will be described briefly here. Potential participants in the study comprised all incident cases of lung cancer aged 20-75 years from the New Zealand Cancer Registry during 2007 and 2008. After notification of a case by the Cancer Registry, both the treating clinician and general practitioner (GP) of the patient were sent a letter explaining the study and asking for consent to contact the patient. For 573 (35%) of the 1,630 notifications nationwide, both the clinician and the GP did not provide consent to contact the patient. Of the 1,057 remaining cases, 116 were not eligible (e.g., never worked in New Zealand, or lung cancer was not the primary cancer). A further 197 patients were deceased and could not be interviewed. From the 744 remaining cases, 283 (38%) declined to participate and 458 were interviewed. Thus, if the patients known to be ineligible for the study and the deceased patients are excluded, the response rate was approximately 53%.

The controls were recruited from the New Zealand Electoral Roll in two time periods, 2003 and in 2008. The former group of 473 controls had been recruited for the previously published studies of NHL, bladder cancer and leukaemia[59-61], whereas 323 controls were newly recruited for the current study during 2008. Controls were frequency matched for the age distribution of registration for these four cancer types in

the New Zealand Cancer Registry. In total, a letter of invitation was sent to 2,000 individuals, of which 122 were returned to sender and thus considered ineligible. Of the remaining 1,878 individuals, contact could not be established for 744 (40%). Their addresses were subsequently compared with the most recent Electoral Rolls. Of the 744 non-responders, 104 did not appear, or appeared with another address, on the new Electoral Roll and were thus considered ineligible. Of the 1,134 for whom contact could be established, 119 were ineligible because of other reasons (e.g., never worked in New Zealand). Of the remaining 1,015 potential controls, 219 (22%) declined to participate, and 796 were interviewed. Thus, if those known to be ineligible for the study are excluded, the response rate in the controls was approximately 48%.

All cases and 364 new controls (46%) were interviewed on the phone, while the other 432 controls were interviewed face-to-face (this was the standard approach in the previous studies for which these controls were initially interviewed). All the interviews were conducted by trained interviewers. The questionnaire included demographic details, a full occupational history and information on potential confounders. Each job held since leaving school was listed, including the start date, the date of termination, the department, the job title, and the name, location and activity of the employer. Then, for each job with a minimum duration of 12 months, more details were asked, including a task description, use of machines and materials, self-reported exposures, workplace ventilation, and use of protective equipment.

Each job was coded according to the 1999 New Zealand Standard Classification of Occupations (NZSCO 1999)[84] (hereafter referred to as the occupational code) and the Australian and New Zealand Standard Industrial Classification (New Zealand use

version 1996)[85] (hereafter referred to as the industry code). These two classifications, based on five and seven digits, respectively, increase the specificity of each occupation/industry with the number of digits. For example, NZCSO code 6 covers ‘‘agriculture and fishery workers’’; code 61 covers ‘‘market oriented agriculture and fishery workers’’; code 611 covers ‘‘market farmers and crop growers.’’ The occupational code was based on the full job and task description, rather than on the occupational title alone, to ensure that the code covered the actual tasks of each job. The industry code was based on the activity of the employer. All coding was done blind to the case-control status of the participants.

Before the data analyses were conducted, two broad lists of *a priori* high-risk occupations and industries were constructed. The selection was based on the international literature and particularly on previous reviews which have listed occupations and industries known[86] and suspected[87] to entail exposure to lung carcinogens[88]. Unconditional regression (SAS V9.1) was used to estimate the odds ratio (OR) and its 95% confidence interval (95% CI), for ever being employed in a certain occupation/industry, compared to never being employed in that occupation/industry. While ORs were calculated for all occupational codes and all industry codes, only those for the 436 occupations and 391 industries in which 5 or more study subjects had ever worked were presented here. ORs were adjusted for age (5-year age-groups), gender, Maōri ethnicity, and smoking status (never, ex, current). The study subjects who reported having stopped smoking less than 2 years before the interview were considered as current smokers. Logistic regression models were also adjusted for socioeconomic status, using the occupational class of the longest held occupation. Occupational class was determined according to the New Zealand Socio-Economic Index (NZSEI), a classification of New Zealand occupations based on

average levels of income and education in national census data[89]. Internal analyses were also conducted to establish whether the duration of employment in a certain occupation or industry was associated with an increased risk. A categorical variable for the duration of each job (<2, 2–10, and>10 years) was created and a test for trend for duration of employment was performed by fitting this categorical variable as a continuous variable in the model.

Semi-Bayes Adjustment

Because of the large number of occupations and industries being considered, this type of study involves multiple comparisons and therefore carries the risk that some of the statistically significant findings are due to chance. As in previous case-control studies in the series[59-61], a semi-Bayes (SB) approach was then used to determine which of the findings were the most robust[48]. The basic idea of SB adjustment for multiple comparisons is that the observed variation of the estimated ORs around their geometric mean will be larger than the variation of the true (but unknown) ORs. The SB method specifies an *a priori* value for the variation of the true ORs; this *a priori* value is then used to adjust the observed ORs[43]. The adjustment consists in shrinking outlying estimates towards the overall mean of the observed estimates. The larger the individual variance of the estimates, the stronger is the shrinkage, that is, the shrinkage is stronger for less reliable estimates based on small numbers. Typical applications in which SB adjustments are a useful addition to traditional methods of adjustment for multiple comparisons include occupational case-control studies, such as the current study, where many relative risks are estimated with few or no *a priori* beliefs about which associations might be causal[43].

SB estimates were calculated using R (free software for statistical computing and graphics)[90]. The inputs for SB adjustments were the maximum likelihood estimates of beta (logOR), resulting from the separate multiple logistic regressions for each occupation and industry. The variance of the true logOR was assumed to be 0.25. Assuming a normal distribution of the logORs, this choice implies that the true ORs are within a sevenfold range of each other[48].

Occupations and industries were divided into groups according to the number of digits of the associated codes, and the shrinkage was performed within these groups. For those occupations (or industries) which were not considered *a priori* to be of increased risk for lung cancer, estimates were shrunk towards the mean for all such occupations (or industries). Similarly, for those occupations (or industries) which were considered to be of *a priori* increased risk for lung cancer, estimates were shrunk towards the mean for all such occupations (or industries).

The findings for all occupations and industries, both before and after SB adjustment, will be made available on web-based tables. Here we report the findings (both before and after SB adjustment) for a priori high-risk occupations and industries and for other occupations and industries that showed statistically significant elevated or decreased risks in the current analyses.

Ethical approval for the study was granted by the Multiregion Ethics Committee (AKL/99/172), and all participants gave informed consent to be interviewed.

Results

The study included 458 interviews with lung cancer cases, and 796 interviews with general population controls. Of these, one case and four controls were excluded due to missing values in key variables, leaving 457 cases and 792 controls available for analysis (Table III.1). Cases were 50% male and controls were 54% male, with a mean age of 60.9 years in cases and 61.5 years in controls. ‘‘Ever smoking’’ was much more frequent among the cases (89%) than among the controls (52%) (OR 7.51, 95% CI 5.41-10.43), with ORs of 14.45 (95% CI 9.10-22.93) for current smokers and 6.62 (95% CI 4.66-9.41) for ex-smokers. Maōri ethnicity was reported by 79 cases (17%) and 22 controls (3%). The occupational class distribution was similar for cases and controls. Logistic regression models were nevertheless adjusted for occupational class in order to be consistent with the previous studies of non-Hodgkin’s lymphoma, bladder cancer and leukaemia[59-61].

Table III.1. Characteristics of the study participants

	Cases		Controls	
	N	**%**	**N**	**%**
Total	457	100.0	792	100.0
Gender				
Men	227	49.7	431	54.4
Women	230	50.3	361	45.6
Age at interview (year)				
20-50	43	9.4	81	10.2
51-60	118	25.8	184	23.2
61-70	284	62.1	424	53.5
>=71	12	2.6	103	13.0
Smoking				
Never	49	10.7	370	46.7
Ex	291	63.7	356	44.9
Current	115	25.2	52	6.6
Missing	2	0.4	14	1.8
Ethnicity				
Maori	79	17.3	22	2.8
Non-Maori	378	82.7	770	97.2
NZSEI				
Class 1 (75-90) highest	7	1.5	33	4.2
Class 2 (60-75)	52	11.4	94	11.9
Class 3 (50-60)	67	14.7	119	15.0
Class 4 (40-50)	114	24.9	164	20.7
Class 5 (30-40)	116	25.4	208	26.3
Class 6 (10-30) lowest	101	22.1	174	22.0

Tables III.2 and III.3 list the findings for the *a priori* high-risk occupations and industries, whereas Table III.4 shows the statistically significant findings for other occupations and industries. Each table shows the findings both before and after SB adjustment, and all analyses are adjusted for the variables specified above. The findings discussed below focus on these high-risk occupations and industries (Tables III.2 and III.3), but also include relevant findings (Table III.4) for other occupations and industries which were not listed as *a priori* at risk.

Wood Workers

Employment as a wood-processing and papermaking plant operator was associated with an increased risk of lung cancer (OR 3.60, 95% CI 0.96-13.48 (not shown in tables)). An increased risk was also observed for sawmill, wood panel and related wood-processing plant operators. In particular, a statistically significant association was found for timber processing machine operators (OR 4.63, 95% CI 1.05-20.29) with a positive relationship between duration of employment and the risk of lung cancer (ORs of 1.11, 4.95, and 14.11 for employment for <2, 2-10, and >10 years, respectively, P-value for linear trend P=0.03 (not shown in tables)). An elevated risk was also observed for loggers (OR 4.67, 95% CI 0.81-27.03) and a statistically significant increased risk was found for men ever employed in the log sawmilling and timber dressing industry (OR 2.85, 95% CI 1.17-6.95 (not shown in tables)). No association with lung cancer was found for carpenters and joiners (OR 1.07, 95% CI 0.61-1.88). Occupations were also classified according to their level of exposure to wood dust using the New Zealand Job Exposure Matrix[91]. Occupations in which at least half of the workers were exposed to average wood dust levels in excess of 0.5 mg/m3 were associated with a slightly elevated risk of lung cancer when compared with those in which wood dust exposure did not occur (OR 1.44, 95% CI 0.84-2.46).

Metal Workers

A statistically significant increased risk of lung cancer was observed for metal and mineral products processing machine operators (OR 4.10, 95% CI 1.37-12.32). Welders and flame-cutters had an increased risk (OR 2.50, 95% CI 0.86-7.25) without a clear association with the duration of employment, and ever being employed in metal ore mining was also associated with an elevated risk (OR 9.92, 95% CI 0.90-109.83 (not

shown in tables)). There was little evidence of an increased risk among other metal workers.

Meat Workers

Statistically significant increased risks of lung cancer were observed for butchers (OR 8.77, 95% CI 1.06-72.55) and for meat and fish processing machine operators (OR2.17, 95% CI 1.22-3.88). Within the last category, fish processing workers had an elevated risk (OR 7.73, 95% CI 0.98-61.13 (not shown in tables)). A statistically significant association was found between duration of employment as a meat and fish processing machine operator and the risk of lung cancer (ORs of 1.95, 3.18, and 1.69 for employment for <2, 2-10, and >10 years, respectively, P=0.02 (not shown in tables)). Workers involved in seafood processing also presented a statistically significant increased risk of lung cancer (OR 4.45, 95% CI 1.02-19.37).

Textile Workers

An increased risk of lung cancer was found for textile products machine operators (OR 1.55, 95% CI 0.97-2.47 (not shown in tables)).Within this category, textile bleaching, dyeing and cleaning machine operators had a statistically significant increased risk of lung cancer (OR 2.35, 95% CI 1.03-5.39); and within this group, pressers were associated with a high lung cancer risk (OR 5.74, 95% CI 0.96-34.42). A duration-response association was found for textile products machine operators (ORs of 1.23, 1.75 and 3.24 for employment for <2, 2-10, and >10 years, respectively, P<0.01 (not shown in tables)), especially for textile bleaching, dyeing and cleaning machine operators (ORs of 1.69, 2.54, and 4.19 for employment for <2, 2-10, and >10 years, respectively, P=0.04 (not shown in tables)). Tailors and dressmakers (OR 2.24, 95% CI

0.84-6.01 (not shown in tables)) and launderers and dry-cleaners (OR 2.25, 95% CI 0.80-6.37 (not shown in tables)) also had increased risks of lung cancer. An elevated risk was observed for individuals employed in the textile product manufacturing industry (OR 1.89, 95% CI 0.88-4.10 (not shown in tables)), and this increased with duration of employment (ORs of 1.15, 1.30, and 11.15 for employment for <2, 2-10, and >10 years, respectively, P=0.02 (not shown in tables)). Women working in textile product manufacturing had a statistically significant increased risk of lung cancer (OR 4.24, 95% CI 1.03-17.44 (not shown in tables)).

Mining Workers

Overall, there was no evidence of an increased risk of lung cancer for subjects involved in mining. There was also no increased risk associated with being employed as a mining and mineral processing plant operator (OR 0.98, 95% CI 0.32-3.03).

Drivers

Working as a driver or mobile machinery operator for at least 10 years was associated with a statistically significant increased risk of lung cancer (OR 2.55, 95% CI 1.31-4.96 (not shown in tables)). For this job category, the risk increased with duration of employment (ORs of 0.49, 0.85, and 2.55 for employment for <2, 2-10, and >10 years, respectively, P=0.08 (not shown in tables)). The same applied for the subcategory of motor vehicle drivers, who had a statistically significant increased risk when they were employed for at least ten years (OR 2.73, 95% CI 1.22-6.07). In particular, statistically significant risks of lung cancer were observed for heavy truck drivers (OR 2.24, 95% CI 1.19-4.21) and for individuals ever employed in the road transport industry (OR 1.78, 95% CI 1.05-3.03) and especially in the road freight transport industry (OR 3.02, 95%

CI 1.45-6.27). A duration-response association was shown for heavy truck drivers (ORs of 1.40, 2.20, and 3.44 for employment for <2, 2-10, and >10 years, respectively, P<0.01 (not shown in tables)) and for subjects employed in road transport (ORs of 0.92, 1.49, and 4.35 for employment for <2, 2-10, and >10 years, respectively, P<0.01 (not shown in tables)) and road freight transport (ORs of 0.97, 3.87, and 6.34 for employment for <2, 2-10, and >10 years, respectively P<0.01 (not shown in tables)). Elevated risks were also found for earthmoving machine operators (OR 5.71, 95% CI 0.87-37.35 (not shown in tables)).

Other *A Priori* High-Risk Occupations

Statistically significant increased risks were found for other rubber and plastics products machine operators (OR 6.34, 95% CI 1.08-37.15) and for individuals employed in the petroleum, coal, chemical and associated product manufacturing industry (OR 1.80, 95% CI 1.11-2.90). Elevated risks were also observed for cement and other minerals processing machine operators (OR 3.73, 95% CI 0.87-15.94), for electric and electronic equipment assemblers (OR 3.61, 95% CI 0.96-13.57) and for machine-tool operators (OR 4.44, 95% CI 0.84-23.63). There was no evidence of an increased risk for the *a priori* occupations of painters and paperhangers, motor mechanics, printing trade workers, chemical processing plant operators, building caretakers and cleaners and for the *a priori* industries of motor vehicle manufacturing, agriculture, printing publishing and recorded media.

Semi-Bayes adjustment of the a priori high-risk occupations and industries

Ever being employed in one or more of the *a priori* high-risk occupations (Table III.2) was associated with a statistically significant increased risk for lung cancer ($OR_{a\ priori}$

$_{\text{occupation}}$ 1.94, 95% CI 1.44-2.61) whereas ever being employed in one or more of the *a priori* high-risk industries (Table III.3) was associated with only a slight increased risk ($OR_{\textit{a priori}\ \text{industry}}$ 1.17, 95% CI 0.89-1.55). This generally resulted in an attenuation of the ORs. Only two of the ORs for the *a priori* high-risk occupations (8323-Heavy truck drivers, odds $\text{ratio}_{\text{Semi-Bayes}}$ (OR_{SB}) 1.97, 95% CI 1.13-3.43 and 83231-Heavy truck or tanker driver OR_{SB} 2.01, 95% CI 1.15-3.50), and one of the ORs for the *a priori* high-risk industries (C25-petroleum, coal, chemical and associated product manufacturing, OR_{SB} 1.66, 95% CI 1.06-2.60) remained significantly elevated after SB adjustment.

Other occupations and industries

Occupations and industries with an observed increased or decreased risk ($P<0.05$), but not considered as *a priori* high risk, are listed in Table III.4.

A number of these occupations and industries showed statistically significant decreased risks for lung cancer. The decreased risks remained statistically significant after SB adjustment for legislators, administrators, and managers (OR_{SB} 0.66, 95% CI 0.49-0.89) and the subcategories of corporate managers, specialized managers, finance and administration managers, for professionals (OR_{SB} 0.31, 95% CI 0.21-0.46) and the subcategories of physical, mathematical and engineering science professionals, teaching professionals, secondary teaching professionals, secondary school teachers, primary and early childhood teaching professionals, primary teaching professionals, primary school teachers, other professionals and business professionals, for other associate professionals (OR_{SB} 0.70, 95% CI 0.50-0.96), for library, mail and related clerks (OR_{SB} 0.66, 95% CI 0.47-0.93) and the subcategories of office clerks and general clerks, for the finance and insurance sector (OR_{SB} 0.60, 95% CI 0.40-0.90), for the government administration and defence sector (OR_{SB} 0.69, 95% CI 0.50-0.94) and the subcategories

of government administration, and for the education sector (OR_{SB} 0.42, 95% CI 0.29-0.61) and the subcategories of school education, primary education, and secondary education. Statistically significant increased risks also appeared among several occupations and industries which were not mentioned above. The associations persisted after SB adjustment for nursing associate professionals (OR_{SB} 2.58, 95% CI 1.29-5.19), and the subcategory of enrolled nurses, for care givers (OR_{SB} 1.99, 95% CI 1.00-3.95), for plant and machine operators and assemblers (OR_{SB} 1.54, 95% CI 1.16-2.04) and the subcategories of stationary machine operators and assemblers and food and related products processing machine operators, for elementary occupations (OR_{SB} 1.39, 95% CI 1.02-1.89) and the subcategories of labourers and related elementary service workers, labourers, builder's labourers, and for the manufacturing industry (OR_{SB} 1.31, 95% CI 1.01-1.70) and for the car retailing industry (OR_{SB} 2.00, 95% CI 1.06-3.80).

Table III.2. Odds ratios (OR) and 95% CIs for a priori high risk occupations

A priori high risk occupation for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95% CI	Global mean[a]
5151-Fire Fighters	3/5	0.76	0.17-3.45	1.20	0.51-2.83	1.45
61122-Grape Grower and/or Wine Maker, Worker	4/2	4.39	0.68-28.51	1.93	0.74-5.05	1.47
61311-Logger	4/3	4.67*	0.81-27.03	2.01	0.76-5.28	1.47
7112-Carpenters and Joiners	25/50	1.07	0.61-1.88	1.16	0.70-1.91	1.45
7124-Painters and Paperhangers	10/16	0.88	0.37-2.11	1.10	0.55-2.18	1.45
722-Blacksmiths, Toolmakers and Related Workers	7/8	1.79	0.58-5.50	1.51	0.71-3.21	1.33
72312-Motor Mechanic	17/25	1.22	0.59-2.51	1.32	0.73-2.39	1.47
733-Printing Trades Workers	9/17	0.87	0.36-2.13	1.06	0.53-2.11	1.33
73321-Bookbinder	2/4	0.90	0.14-5.56	1.37	0.56-3.33	1.47
7411-Butchers	8/1	8.77**	1.06-72.55	2.00	0.70-5.73	1.45
742-Cabinet Makers and Related Workers	7/7	2.11	0.63-7.09	1.59	0.72-3.54	1.33
811-Mining and Mineral Processing Plant Operators	6/11	0.98	0.32-3.03	1.16	0.55-2.48	1.33
8111-Mining Plant Operators	5/6	1.93	0.44-8.43	1.58	0.70-3.61	1.45
81111-Quarry and Mine Worker	5/5	1.97	0.44-8.78	1.66	0.73-3.79	1.47
8113-Drillers	1/6	0.25	0.03-2.17	1.08	0.38-3.06	1.45
8122-Metal Casters	3/2	1.82	0.24-13.81	1.52	0.63-3.68	1.45
81231-Welder and Flame-Cutter	12/7	2.50*	0.86-7.25	1.92	0.90-4.10	1.47
813-Glass and Ceramics Kiln and Related Plant Operators	7/9	1.60	0.51-5.03	1.44	0.68-3.04	1.33
8131-Glass and Ceramics Kiln Operators	4/7	1.15	0.27-4.89	1.35	0.60-3.05	1.45
81312-Clay Product Plant Operator	2/3	1.26	0.13-12.40	1.49	0.61-3.69	1.47

A priori high risk occupation for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95% CI	Global mean[a]
8132-Other Glass and Ceramics Workers	3/2	2.85	0.40-20.47	1.66	0.67-4.11	1.45
8141-Sawmill, Wood Panel and Related Wood-Processing Plant Operators	8/5	4.63**	1.05-20.29	2.07	0.82-5.23	1.45
815-Chemical Processing Plant Operators	8/9	1.43	0.47-4.32	1.37	0.66-2.86	1.33
8211-Machine-Tool Operators	8/2	4.44*	0.84-23.63	1.93	0.76-4.93	1.45
8212-Cement and Other Minerals Processing Machine Operators	7/3	3.73*	0.87-15.94	1.95	0.80-4.75	1.45
8222-Metal Finishers, Platers and Coaters	5/6	0.87	0.25-3.03	1.20	0.54-2.65	1.45
82221-Electroplater	2/4	0.63	0.11-3.55	1.24	0.49-3.14	1.47
8231-Tyre Production Machine Operators	4/2	2.45	0.36-16.65	1.62	0.67-3.94	1.45
8232-Other Rubber and Plastics Products Machine Operators	8/2	6.34**	1.08-37.15	2.05	0.76-5.59	1.45
8261-Spinning and Winding Machine Operators	3/3	1.77	0.29-11.00	1.52	0.64-3.61	1.45
8262-Weaving and Knitting Machine Operators	7/9	1.47	0.49-4.41	1.46	0.70-3.04	1.45
82624-Knitter, Knitting Machinist	2/3	1.63	0.23-11.71	1.56	0.65-3.75	1.47
82641-Launderer	9/5	2.29	0.60-8.73	1.77	0.79-3.99	1.47
82643-Dry-Cleaner	3/4	1.02	0.19-5.38	1.39	0.59-3.28	1.47
82644-Presser	6/2	5.74*	0.96-34.42	2.09	0.76-5.73	1.47
82651-Fibre Preparer	2/3	0.75	0.09-6.19	1.36	0.54-3.43	1.47
82812-Tanner, Splitter and Dyer	3/2	1.15	0.16-8.26	1.46	0.60-3.52	1.47
82922-Electric and Electronic Equipment Assembler	7/5	3.61*	0.96-13.57	2.08	0.87-5.02	1.47
83212-Light Truck or Van Driver	5/11	0.55	0.17-1.81	1.02	0.41-2.49	1.47
8322-Bus Drivers	5/12	0.46	0.13-1.68	0.96	0.39-2.36	1.45
8323-Heavy Truck Drivers	31/26	2.24**	1.19-4.21	1.97	1.13-3.43	1.45

A priori high risk occupation for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95% CI	Global mean[a]
83231-Heavy Truck or Tanker Driver	31/26	2.24**	1.19-4.21	2.01	1.15-3.50	1.47
8331-Motorised Farm Machinery Operators	5/10	0.72	0.22-2.36	1.09	0.49-2.46	1.45
8332-Earthmoving and Related Machinery Operators	9/13	1.01	0.37-2.76	1.22	0.59-2.49	1.45
83325-Roading and/or Paving Machine Operator	1/9	0.15	0.01-1.48	1.08	0.32-3.62	1.47
84117-Roofer	1/4	0.91	0.09-9.47	1.43	0.57-3.58	1.47
911-Building Caretakers and Cleaners	50/50	1.23	0.75-2.00	1.25	0.80-1.93	1.33

** p<0.05 *p<0.1 [a] Weighted geometric mean of the ORs (see Chapter II for calculation details)

Table III.3. Odds Ratios (OR) and 95% CIs for *a priori* high risk industries

A priori high risk industry for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95% CI	Global mean[a]
A01-Agriculture	99/183	1.01	0.73-1.40	1.03	0.75-1.40	1.19
C2111-Meat Processing	39/41	1.17	0.68-2.00	1.19	0.74-1.92	1.29
C226-Leather and Leather Product Manufacturing	5/7	0.96	0.25-3.70	0.84	0.37-1.91	0.79
C24-Printing, Publishing and Recorded Media	21/39	0.92	0.50-1.70	0.99	0.58-1.68	1.19
C25-Petroleum, Coal, Chemical and Associated Product Manufacturing	51/52	1.80**	1.11-2.90	1.66*	1.06-2.60	1.19
C2551-Rubber Tyre Manufacturing	4/3	1.60	0.30-8.46	1.36	0.58-3.19	1.29
C261-Glass and Glass Product Manufacturing	4/11	0.51	0.14-1.83	0.67	0.28-1.61	0.79
C264-Non-Metallic Mineral Product Manufacturing not elsewhere classified	3/4	1.67	0.33-8.39	0.96	0.33-2.80	0.79
C2711-Basic Iron and Steel Manufacturing	5/5	1.47	0.38-5.75	1.35	0.61-3.00	1.29
C272-Basic Non-Ferrous Metal Manufacturing	2/6	0.62	0.11-3.59	0.74	0.31-1.79	0.79
C2811-Motor Vehicle Manufacturing	13/16	1.19	0.51-2.74	1.23	0.65-2.32	1.29
C2821-Shipbuilding	3/2	2.95	0.44-19.60	1.53	0.58-4.05	1.29
C2859-Electrical Equipment Manufacturing not elsewhere classified	8/10	1.30	0.46-3.68	1.29	0.63-2.64	1.29
D362-Gas Supply	2/4	0.57	0.07-4.57	0.74	0.30-1.86	0.79

** p<0.05 *p<0.1 [a] Weighted geometric mean of the ORs (see Chapter II for calculation details)

Table III.4. Odds Ratios (OR) and 95% CIs for not *a priori* high risk occupations and industries (p<0.05) (excluding the *a priori* high risk occupations listed in tables III.2 and III.3)

Not *a priori* high risk occupation and industry for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95%CI	Global mean[a]
Occupations-reduced risk						
1-Legislators, Administrators and Managers	102/239	0.64**	0.47-0.87	0.66	0.49-0.89	0.93
12-Corporate Managers	101/235	0.65**	0.48-0.89	0.67	0.50-0.90	0.92
122-Specialised Managers	88/206	0.62**	0.45-0.85	0.65	0.48-0.88	1.00
1221-Production and Operation Managers	15/56	0.50**	0.27-0.95	0.61	0.35-1.04	0.95
1222-Finance and Administration Managers	9/40	0.32**	0.15-0.70	0.49	0.26-0.92	0.95
12222-Administration Manager	3/30	0.17**	0.05-0.58	0.52	0.23-1.15	1.05
2-Professionals	71/247	0.26**	0.18-0.39	0.31	0.21-0.46	0.93
21-Physical, Mathematical and Engineering Science Professionals	14/53	0.44**	0.23-0.85	0.55	0.32-0.97	0.92
214-Architects, Engineers and Related Professionals	11/36	0.47**	0.22-1.00	0.62	0.34-1.13	1.00
23-Teaching Professionals	25/115	0.25**	0.15-0.43	0.34	0.20-0.56	0.92
231-Tertiary Teaching Professionals	7/31	0.28**	0.11-0.71	0.51	0.25-1.02	1.00
23111-University and Higher Education Lecturer and/or Tutor	7/30	0.29**	0.12-0.72	0.52	0.26-1.05	1.05
232-Secondary Teaching Professionals	5/43	0.20**	0.07-0.55	0.46	0.22-0.96	1.00
23211-Secondary School Teacher	5/43	0.20**	0.07-0.55	0.47	0.22-0.98	1.05
233-Primary and Early Childhood Teaching Professionals	15/64	0.25**	0.13-0.50	0.40	0.22-0.72	1.00
2331-Primary Teaching Professionals	10/55	0.24**	0.11-0.52	0.41	0.21-0.77	0.95
23311-Primary School Teacher	10/55	0.24**	0.11-0.52	0.42	0.22-0.79	1.05
24-Other Professionals	11/67	0.23**	0.11-0.47	0.37	0.20-0.69	0.92

Not *a priori* high risk occupation and industry for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95%CI	Global mean[a]
241-Business Professionals	9/40	0.29**	0.13-0.65	0.48	0.25-0.92	1.00
2411-Accountants	3/24	0.19**	0.05-0.68	0.52	0.23-1.18	0.95
24111-Accountant	3/24	0.19**	0.05-0.68	0.55	0.25-1.24	1.05
33-Other Associate Professionals	77/169	0.67**	0.48-0.95	0.70	0.50-0.96	0.92
4114-Secretaries	14/59	0.47**	0.24-0.90	0.58	0.33-1.02	0.95
41141-Secretary	14/59	0.47**	0.24-0.90	0.60	0.34-1.05	1.05
412-Numerical Clerks	22/58	0.54**	0.31-0.94	0.63	0.38-1.02	1.00
414-Library, Mail and Related Clerks	66/164	0.62**	0.43-0.89	0.66	0.47-0.93	1.00
4144-Office Clerks	53/141	0.57**	0.39-0.85	0.61	0.42-0.89	0.95
41443-General Clerk	51/131	0.58**	0.39-0.87	0.63	0.44-0.92	1.05
4212-Bank Officers	5/32	0.29**	0.11-0.78	0.53	0.26-1.09	0.95
42121-Bank Officer	5/32	0.29**	0.11-0.78	0.55	0.27-1.14	1.05
51233-Waiter	7/23	0.29**	0.11-0.77	0.55	0.27-1.11	1.05
51312-Health Assistant	2/10	0.19**	0.04-1.00	0.67	0.28-1.59	1.05
6122-Mixed Livestock Producers	2/27	0.14**	0.03-0.63	0.54	0.22-1.28	0.95
61221-Mixed Livestock Farmer, Mixed Livestock Farm Worker	2/27	0.14**	0.03-0.63	0.57	0.24-1.35	1.05
Occupations-increased risk						
3118-Draughting Technicians	4/3	8.70**	1.22-61.99	1.48	0.59-3.72	0.95
31181-Draughting Technician	4/3	8.70**	1.22-61.99	1.59	0.64-3.93	1.05
32-Life Science and Health Associate Professionals	32/32	1.89**	1.05-3.37	1.56	0.94-2.60	0.92
323-Nursing Associate Professionals	25/9	5.45**	2.29-12.99	2.58	1.29-5.19	1.00
32311-Enrolled Nurse	25/7	7.95**	3.10-20.42	2.99	1.44-6.22	1.05

Not *a priori* high risk occupation and industry for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95%CI	Global mean[a]
51231-Bartender	15/11	2.98**	1.17-7.61	1.80	0.90-3.60	1.05
51316-Care Giver	20/10	3.47**	1.40-8.59	1.99	1.00-3.95	1.05
61213-Cattle Farmer, Cattle Farm Worker	7/3	5.56**	1.21-25.51	1.70	0.73-3.97	1.05
614-Fishery Workers, Hunters and Trappers	11/8	3.91**	1.24-12.37	1.77	0.82-3.83	1.00
7111-Bricklayers and Stonemasons	7/3	5.65**	1.24-25.83	1.60	0.68-3.81	0.95
71111-Bricklayer and/or Blocklayer	7/3	5.65**	1.24-25.83	1.71	0.73-4.00	1.05
74-Other Craft and Related Trades Workers	29/27	2.12**	1.14-3.93	1.67	0.98-2.85	0.92
8-Plant and Machine Operators and Assemblers	204/233	1.61**	1.20-2.16	1.54	1.16-2.04	0.93
81411-Timber processing Machine operator	8/5	4.63**	1.05-20.29	1.64	0.71-3.79	1.05
82-Stationary Machine Operators and Assemblers	148/150	1.67**	1.22-2.28	1.58	1.16-2.13	0.92
821-Metal and Mineral Products Processing Machine Operators	15/5	4.10**	1.37-12.32	1.86	0.87-3.99	1.00
823-Rubber and Plastics Products Machine Operator	12/4	4.27**	1.16-15.66	1.69	0.75-3.79	1.00
82322-Plastics Machine Operator	8/2	6.34**	1.08-37.15	1.59	0.66-3.84	1.05
8264-Textile Bleaching, Dyeing and Cleaning Machine Operators	20/13	2.35**	1.03-5.39	1.61	0.84-3.09	0.95
827-Food and Related Products Processing Machine Operators	63/46	1.98**	1.23-3.19	1.73	1.12-2.68	1.00
8271-Meat and Fish Processing Machine Operators	45/27	2.17**	1.22-3.88	1.75	1.05-2.93	0.95
8292-Electrical Machinery Assemblers	14/12	2.84**	1.15-6.97	1.72	0.87-3.42	0.95
9-Elementary Occupations (incl Residuals)	128/146	1.45**	1.05-2.00	1.39	1.02-1.89	0.93
91-Labourers and Related Elementary Service Workers	128/146	1.45**	1.05-2.00	1.39	1.02-1.89	0.92
915-Labourers	51/54	1.91**	1.19-3.05	1.69	1.10-2.60	1.00
91512-Builder's Labourer	20/11	3.15**	1.40-7.11	2.01	1.06-3.82	1.05

Not *a priori* high risk occupation and industry for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95%CI	Global mean[a]
Industries-reduced risk						
A0123-Sheep-Beef Cattle Farming	7/31	0.30**	0.12-0.79	0.56	0.28-1.14	1.08
F46-Machinery and Motor Vehicle Wholesaling	11/37	0.43**	0.21-0.91	0.60	0.33-1.10	1.06
F462-Motor Vehicle Wholesaling	1/13	0.09**	0.01-0.73	0.68	0.27-1.70	1.05
G511-Supermarket and Grocery Stores	31/81	0.61**	0.37-0.98	0.67	0.44-1.04	1.05
K-Finance and Insurance	37/107	0.55**	0.36-0.85	0.60	0.40-0.90	0.93
K73-Finance	21/61	0.57**	0.32-0.99	0.66	0.40-1.08	1.06
K75-Services to Finance and Insurance	7/43	0.30**	0.13-0.71	0.52	0.27-1.02	1.06
K752-Services to Insurance	6/33	0.38**	0.15-0.97	0.62	0.31-1.23	1.05
L7842-Accounting Services	7/35	0.41**	0.17-0.98	0.63	0.32-1.23	1.08
M-Government Administration and Defence	87/186	0.67**	0.48-0.92	0.69	0.50-0.94	0.93
M81-Government Administration	62/136	0.63**	0.43-0.92	0.67	0.47-0.96	1.06
M811-Government Administration	57/129	0.61**	0.42-0.90	0.66	0.46-0.95	1.05
M8113-Local Government Administration	20/49	0.49**	0.27-0.91	0.61	0.36-1.04	1.08
N-Education	56/192	0.37**	0.25-0.55	0.42	0.29-0.61	0.93
N84-Education	56/191	0.38**	0.26-0.55	0.43	0.30-0.63	1.06
N842-School Education	35/134	0.34**	0.22-0.54	0.42	0.27-0.64	1.05
N8421-Primary Education	18/66	0.37**	0.20-0.68	0.50	0.29-0.85	1.08
N8422-Secondary Education	16/68	0.39**	0.20-0.73	0.52	0.30-0.91	1.08
N843-Post School Education	15/53	0.46**	0.24-0.91	0.60	0.34-1.06	1.05
N8431-Higher Education	5/32	0.26**	0.09-0.74	0.55	0.26-1.15	1.08
N844-Other Education	1/57	0.03**	0.00-0.21	0.52	0.20-1.34	1.05

Not *a priori* high risk occupation and industry for lung cancer		Not semi-Bayes adjusted		Semi-Bayes adjusted		
	cases/controls (n)	OR	95% CI	OR	95%CI	Global mean[a]
Industries-increased risk						
A04-Commercial Fishing	12/6	5.60**	1.53-20.44	1.94	0.86-4.40	1.06
A041-Marine Fishing	9/3	6.33**	1.18-34.11	1.65	0.69-3.94	1.05
C-Manufacturing	255/354	1.34**	1.02-1.77	1.31	1.01-1.70	0.93
C2173-Seafood Processing	8/4	4.45**	1.02-19.37	1.67	0.72-3.84	1.08
C256-Plastic Product Manufacturing	17/10	3.11**	1.19-8.16	1.82	0.91-3.67	1.05
E4243-Tiling and Carpeting Services	8/2	9.53**	1.81-50.06	1.89	0.78-4.57	1.08
G531-Motor Vehicle Retailing	18/15	2.60**	1.19-5.70	1.83	0.98-3.40	1.05
G5311-Car Retailing	18/13	3.08**	1.36-6.94	2.00	1.06-3.80	1.08
I61-Road Transport	38/42	1.78**	1.05-3.03	1.58	0.99-2.54	1.06
I611-Road Freight Transport	25/17	3.02**	1.45-6.27	2.07	1.14-3.77	1.05
I6623-Port Operators	8/3	4.95**	1.08-22.66	1.68	0.72-3.92	1.08
K7422-General Insurance	12/9	2.88**	1.01-8.18	1.71	0.82-3.53	1.08
O8612-Psychiatric Hospitals	14/3	8.90**	2.07-38.25	2.08	0.89-4.88	1.08

** $p<0.05$ [a] Weighted geometric mean of the ORs (see Chapter II for calculation details)

Discussion

This study of 457 incident lung cancer cases diagnosed in New Zealand during 2007 and 2008 and 792 population controls, aimed to identify occupations that entail an elevated risk for lung cancer. After adjustment for age, gender, smoking status, Maōri ethnicity and occupational status, the analyses showed that wood workers, metal workers, textile workers, meat workers and drivers had increased risks for lung cancer, and that several other occupations and industries (including nursing associate professionals, labourers) also had an increased risk for lung cancer in the New Zealand population.

Before discussing the detailed study findings, the strengths and limitations of this study should be considered. The use of the New Zealand Cancer Registry and of the Electoral Roll to identify cases and to sample controls presented several advantages. The Cancer Registry covers all primary malignant cancers diagnosed in New Zealand and the Electoral Roll records virtually all New Zealand citizens and permanent residents aged 18 years and older (registration on the Electoral Roll is compulsory). They therefore provide reliable sources for the selection of cases and controls. The low response rate remains one of the most important limitations in this series of studies. However, the Electoral Roll records occupation, thus making it possible to compare the occupations of participating and non-participating controls. In previous studies, although there was little evidence of a systematic response bias, we have found that participation rates were lower for the lowest occupational class. We therefore adjusted for occupational class in the analyses, as we have done in the current study, but this adjustment has made little difference to the study findings. In addition, we have adjusted for tobacco smoking in the analyses, and this has also made little difference to the study findings.

Although a strength of a study such as this is the efficiency of evaluating potential associations with many occupational categories, it also introduces the problem of multiple comparisons and the possibility that some statistically significant findings may be due to chance. For this reason, we used the SB adjustment method, which helped us to assess which findings were most robust. The validity of these SB adjustments depends on the assumption that the *a priori* identified occupations have associations that are ‘‘exchangeable’’ which may not always be the case, but this is unlikely to have appreciably affected the results. A further limitation of the current study is that the controls were recruited during two time periods, 2003 and 2008, whereas the cases were only recruited during the latter time period. However, the difference in the time periods was only 5 years, and there was little change in New Zealand employment patterns, or unemployment rates, during this time[92]. Finally, we were able to adjust for smoking status, but not smoking duration; however, information on smoking status has previously been found to allow satisfactory control of confounding by smoking[93].

Wood Workers

Wood dust created by wood processing has been classified as a carcinogen by the International Agency for Research on Cancer[94], particularly for sinus and nasal cancer. Some studies have also reported that wood dust exposure may be a risk factor for lung cancer[95-97]. In New Zealand, increased risks of lung cancer have been previously reported for sawmillers and carpenters in a series of case-control studies[83]. In the current study, statistically significant elevated risks of lung cancer were observed for sawmill, wood panel and related wood-processing plant operators and especially for timber processing machine operators, but there was no evidence of an increased risk for carpenters and joiners. For timber processing machine operators, the risk of lung cancer

went up with the duration of employment. Finally, a slightly increased risk was found among workers with regular exposure to wood dust at levels above 0.5 mg/m^3.

Metal Workers

Increased risks of lung cancer have already been reported for occupations in the metal industry[98-103]. These occupations can entail exposure to asbestos, a well-known lung carcinogen, but also to potentially carcinogenic metal fumes and dust (e.g., arsenic, chromium)[104, 105]. In the current study, the risk of lung cancer was increased only for welders and flame cutters. This association has been observed previously in a number of studies[44, 106-109].

Meat Workers

Elevated risks of lung cancer have been reported for butchers and meat workers in several studies[82, 110-112]. In New Zealand, slaughtering of sheep and cattle (in the ‘‘freezing industry’’) is a major economic activity, and workers are exposed to blood, faeces, urine, and other biological exposures[61]. A New Zealand study[113] observed an excess risk of lung cancer (RR=1.79) in a cohort of three meat processing plants. In the current study (which had no overlap of cases with the previous studies), statistically significant increased risks were observed for butchers, for meat and fish processing machine operators and for individuals employed in seafood processing.

Textile Workers

An elevated risk of lung cancer was observed for workers in the textile product manufacturing industry, for textile products machine operators, and particularly for textile bleaching, dyeing and cleaning machine operators and for pressers. Increased

risks of lung cancer have already been found for pressers[114] and more generally for workers engaged in dry cleaning and laundering[44, 115-117]. These occupations entail exposure to organic solvents and textile dyes and some of these substances have been demonstrated to be carcinogenic[118]. Decreased risks of lung cancer have been reported for cotton and wool textile workers in several studies[119-123]. However, the production of some textiles may also involve exposure to asbestos[124-126], a recognized lung carcinogen.

Drivers

In this study, heavy truck drivers and workers employed in road transport and in road freight transport had a statistically significant increased risk of lung cancer. The International Agency for Research on Cancer has defined diesel and gasoline vehicle exhaust as, respectively, probably and possibly carcinogenic to humans[127]. Several studies have reported elevated risks of lung cancer for motor vehicle drivers[98, 128-132].

Other Occupations and Industries

In the current study, elevated risks of lung cancer were observed for life science and health associate professionals and in particular for nurses. Health care workers have already been reported to be at risk for lung cancer in previous studies[133, 134], and exposure to ionizing radiation is one possible explanation. Employment as a labourer and particularly as a builder's labourer was also associated with an elevated risk in this study. Construction labourers are potentially exposed to asbestos, organic solvents and dust, and they have been shown in previous studies to have an excess risk of lung cancer[98, 135, 136]. Several statistically significant decreased risks of lung cancer

were also found for occupations held by white-collar workers and the corresponding industries.

Conclusions

These findings of this study indicate increased lung cancer risks associated with certain occupations and industries in New Zealand, including wood workers, metal workers, meat workers, textile workers, and drivers. Further analyses should be conducted to determine which particular carcinogenic exposures occur for individuals holding these occupations and how the intensity of these exposures affects the risk of lung cancer.

Acknowledgements

This project was funded by the Health Research Council of New Zealand, by the New Zealand Department of Labour, by Lottery Health Research, by the Cancer Society of New Zealand, and by the Accident Compensation Corporation (ACC). Views and/or conclusions in this article are those of Massey University and may not reflect the position of ACC. The Centre for Public Health Research is supported by a Programme Grant from the Health Research Council. We thank Pam Miley-Terry, Joy Stubbs, Nicky Curran, and Heather Duckett. We also thank the staff of the New Zealand Cancer Registry at the New Zealand Health Information Service for collecting and making available information on cancer registrations. We also thank Miria Hudson for her assistance.

CHAPTER IV

Occupation and risk of upper aerodigestive tract cancer: the ARCAGE study

Lorenzo Richiardi, Marine Corbin, Manuela Marron, Wolfgang Ahrens, Hermann Pohlabeln, Pagona Lagiou, Ploumitsa Minaki, Antonio Agudo, Xavier Castellsague, Alena Slamova, Miriam Schejbalova, Kristina Kjaerheim, Luigi Barzan, Renato Talamini, Gary J. Macfarlane, Tatiana V. Macfarlane, Cristina Canova, Lorenzo Simonato, David I. Conway, Patricia A. McKinney, Linda Sneddon, Peter Thomson, Ariana Znaor, Claire M. Healy, Bernard E. McCartan, Simone Benhamou, Christine Bouchardy, Mia Hashibe, Paul Brennan and Franco Merletti

We investigated the association between occupational history and upper aerodigestive tract (UADT) cancer risk in the ARCAGE European case-control study. The study included 1,851 patients with incident cancer of the oral cavity, oropharynx, hypopharynx, larynx or oesophagus and 1,949 controls. We estimated odds ratios (OR) and 95% confidence intervals (CI) for ever employment in 283 occupations and 172 industries, adjusting for smoking and alcohol. Men (1,457 cases) and women (394 cases) were analysed separately and we incorporated a semi-Bayes adjustment approach for multiple comparisons. Among men, we found increased risks for occupational categories previously reported to be associated with at least one type of UADT cancer, including painters (OR=1.74, 95% CI: 1.01-3.00), bricklayers (1.58, 1.05-2.37), workers employed in the erection of roofs and frames (2.62, 1.08-6.36), reinforced concreters (3.46, 1.11-10.8), dockers (2.91, 1.05-8.05) and workers employed in the construction of roads (3.03, 1.23-7.46), general construction of buildings (1.44, 1.12-1.85) and cargo handling (2.60, 1.17-5.75). With the exception of the first three categories, risks both increased when restricting to long

duration of employment and remained elevated after semi-Bayes adjustment. Increased risks were also found for loggers (3.56, 1.20-10.5) and cattle and dairy farming (3.60, 1.15-11.2). Among women, there was no clear evidence of increased risks of UADT cancer in association with occupations or industrial activities. This study provides evidence of an association between some occupational categories and UADT cancer risk among men. The most consistent findings, also supported by previous studies, were obtained for specific workers employed in the construction industry.

Key words: occupational exposures, upper aerodigestive tract cancer, case-control study

Chapter based on International Journal of Cancer 2012; 130(10), 2397–2406

Alcohol drinking and tobacco smoking are the two main risk factors for cancers of the upper aerodigestive tract (UADT), which group together tumours originating in the oral cavity, pharynx, larynx and oesophagus[137-139]. These two exposures may explain up to 75% of all UADT cancer cases[139]. Diet[140], human papillomavirus (HPV) infection[141], low socioeconomic status[142] and genetic susceptibility[143] have all been indicated as other potential risk factors.

A number of case-control studies have investigated occupational exposures in relation to the risk of UADT cancer[144-159]. Most of these studies had a limited sample size and focused on laryngeal cancer only. Increased risks have been repeatedly found for a number of occupations, including painters[145-148, 160], specific categories of construction workers,[145-150, 152, 153] metal workers[145-150], labourers[145, 147, 149-152, 154, 155], butchers[150, 157] and shoe, leather and textile workers[150, 151, 155, 157, 159-161] as well as for exposure to some specific occupational agents, such as sulphuric acid, asbestos and coal dust[156, 162, 163].

We used data from a large multicentre case-control study recently conducted in 14 centres throughout Europe, in which a detailed occupational history was collected using a standardized questionnaire, to further investigate the role for occupational factors in UADT cancer aetiology.

Methods

Study design and exposure information

ARCAGE (Alcohol Related Cancers and Genetic Susceptibility in Europe) is a European multicentre case-control study on UADT cancer carried out between 2000 and 2005 in 14 centres in 10 European countries, including Czech Republic, Croatia, France (in which recruitment was conducted between 1987 and 1992), Germany, Greece, Ireland, Italy, Norway, Spain and the UK. It was approved by the ethical committee of the coordinating centre (International Agency for Research on Cancer, Lyon, France) and the local ethical committees at each participating centre. A detailed description of the study methods has been published before[164].

The study was hospital-based in most of the countries, with the exception of the three UK centres in which a population-based approach was used. In each centre, cases included all newly diagnosed primary cancers occurring in the oral cavity (ICD-O-3: C00-C06), oropharynx (C09, C10), hypo-pharynx (C12, C13), larynx (C14, C32) or oesophagus (C15) identified by constant monitoring at the hospitals and clinics participating in the study. All cases were histologically confirmed.

Controls were frequency-matched to cases by 5-year age groups, sex and centre. Hospital controls were selected among patients admitted for diseases unrelated to

tobacco or alcohol. Eligible diagnoses included endocrine and metabolic disorders as well as genito-urinary, skin, subcutaneous tissue and musculoskeletal diseases, gastro-intestinal, circulatory, ear, eye and mastoid diseases, nervous system diseases, trauma and plastic surgery patients. The proportion of controls within a specific diagnostic group should not exceed 33% of the total in any centre. In the UK, population controls were randomly selected from a list of individuals registered with the same general practitioner as the corresponding cases. In the Paris centre, the source population was limited to smokers.

We used a face-to-face interview and a standardized questionnaire to obtain information from all study subjects on demographic characteristics, educational level, lifetime smoking, alcohol consumption, diet, medical history, anthropometric measures and full occupational history. For each occupational period that lasted at least 6 months, we recorded the year of beginning and end as well as the descriptions of the job title and the branch of industry. Part-time and seasonal jobs were recorded.

In each country, a trained coder codified the occupational periods blinded to case-control status according to the National Industrial Classification of All Economic Activities (NACE)[165] for branches of industry and the International Standard Classification of Occupations (ISCO)[45] for the job titles. These classifications, based on four and five digits respectively, increase the specificity of each occupation/industry with increasing number of digits. The Paris centre had coded occupational histories differently, namely using 3-digit ISCO codes for the job titles and the ISIC-2 classification for the branches of industries. We therefore excluded the Paris centre from the main analyses, although we carried out a separated analysis based on 3-digit ISCO

codes among the 297 male cases and the 210 male controls from Paris to check for consistency with the results obtained on all other centres.

Statistical analyses

Overall, 1,981 cases and 1,993 controls participated in the study with a response proportion of 82% among cases and 68% among controls. We excluded 77 cases with adenocarcinoma of the oesophagus and 18 subjects with *in-situ* carcinoma. Moreover, we excluded all case and control subjects with missing values in smoking, alcohol consumption and/or educational level (27 cases and 32 controls). Another 8 cases and 12 controls were excluded because they had no information on their occupational history, thus leaving 1,851 cases and 1,949 controls for the present analyses.

We carried out analyses in men (1,457 cases and 1,425 controls) and women (394 cases and 524 controls) separately and used multivariable logistic regression to estimate odds ratios of UADT cancer, with corresponding 95% confidence intervals (95% CI), for ever compared with never employment in each occupational or industrial category. We considered a lag time of 10 years, thus exposures occurring in the last ten years before the interview were not considered, and analysed only categories including at least 10 exposed subjects. All models included centre, age, cigarette smoking and alcohol consumption. These variables were categorized as reported in Table IV.1. As previously suggested,[166] we estimated each odds ratio (OR) in models with and without adjustment for attained educational level. All UADT cancer cases were grouped together in the main analyses while we conducted secondary analyses considering the three main subsites separately (mouth and oropharynx, hypopharynx and larynx, oesophagus). We also carried out the main analyses for men, including centre as a random effect in the models. Since the estimates obtained from the fixed effects models

and from the random effects models were very similar, we only reported those obtained from the fixed effects models. Analyses were conducted using the software SAS, version 9.

Since we considered a large number of occupations and industries, we also applied a semi-Bayes (SB) approach[17, 43], using R software, to identify the most robust estimates. We assumed a variance of the true Log ORs of 0.25 and shrunk the estimates for each category towards the overall mean, for the industries, and towards a group mean, for the occupations, where we used two groups, namely blue-collar worker and white-collar worker occupations.

Results

Characteristics of cases and controls are reported in Table IV.1. Tobacco smoking and alcohol consumption were higher among cases than controls both in men and women. Cases had a lower educational level, while the mean number of job periods was similar between cases and controls. Among men and women combined, 49% of the UADT cancers occurred in the oral cavity/oropharynx, 37% in the hypopharynx/larynx, 9% in the oesophagus and 5% were classified as overlapping cancers.

Table IV.1. Selected characteristics of cases and controls

CHARACTERISTICS	MEN				WOMEN			
	Cases (N=1457)		Controls (N=1425)		Cases (N=394)		Controls (N=524)	
	N	%	N	%	N	%	N	%
Center								
Prague	142	9.8	145	10.2	26	6.6	22	4.2
Bremen	229	15.7	261	18.3	51	12.9	64	12.2
Athens	181	12.4	142	10.0	43	10.9	51	9.7
Aviano	118	8.1	118	8.3	32	8.1	33	6.3
Padova	106	7.3	97	6.8	26	6.6	33	6.3
Turin	116	8.0	144	10.1	43	10.9	52	9.9
Dublin	22	1.5	6	0.4	10	2.5	12	2.3
Oslo	109	7.5	109	7.7	42	10.7	70	13.4
Glasgow	61	4.2	51	3.6	31	7.9	39	7.4
Manchester	104	7.1	122	8.6	42	10.7	62	11.8
Newcastle	59	4.1	94	6.6	14	3.6	18	3.4
Barcelona	165	11.3	99	7.0	25	6.4	59	11.3
Zagreb	45	3.1	37	2.6	9	2.3	9	1.7
Age (years)								
<40	16	1.1	43	3.0	21	5.3	46	8.8
40-44	53	3.6	62	4.4	21	5.3	24	4.6
45-49	144	9.9	121	8.5	25	6.4	36	6.9
50-54	243	16.7	209	14.7	35	8.9	60	11.5
55-59	307	21.1	257	18.0	75	19.0	86	16.4
60-64	245	16.8	216	15.2	57	14.5	61	11.6
65-69	208	14.3	231	16.2	62	15.7	79	15.1
70-74	140	9.6	171	12.0	53	13.5	62	11.8
75-79	101	6.9	115	8.1	45	11.4	70	13.4
Education*								
Low	548	37.6	342	24.0	126	32.0	156	29.8
Medium	837	57.5	926	65.0	249	63.2	330	63.0
High	72	4.9	157	11.0	19	4.8	38	7.2
Smoking – packyears								
Never	84	5.8	407	28.6	116	29.4	299	57.1
<20 packyears	215	14.8	427	30.0	103	26.1	135	25.8
20-39 packyears	498	34.2	350	24.6	117	29.7	62	11.8
>=40 packyears	660	45.3	241	16.9	58	14.7	28	5.3
Alcohol – duration								
Never	37	2.5	96	6.7	79	20.1	151	28.8
1-9 years	20	1.4	34	2.4	16	4.1	26	5.0
10-19 years	51	3.5	69	4.8	36	9.1	48	9.2
20-29 years	209	14.3	171	12.0	72	18.3	78	14.9
30-39 years	466	32.0	396	27.8	94	23.9	108	20.6
40+ years	674	46.3	659	46.3	97	24.6	113	21.6
N. of jobs								
Average	3.7		3.7		3.6		3.5	
Topography			-	-			-	-
Oral/Oropharynx	670	46.0	-	-	245	62.2	-	-
Hypopharynx/Larynx	611	41.9	-	-	75	19.0	-	-
Esophagus	107	7.3	-	-	47	11.9	-	-
Overlapping	69	4.7	-	-	27	6.9	-	-

***** Low=Finished primary school; Medium=Finished other school; High=University degree

Men

Overall, we evaluated 283 occupational categories (3 and 5 digits), 17 of which were associated with UADT cancer risk with a p-value below 0.05 in smoking- and alcohol-adjusted analyses. Out of these, 10 categories were associated with an increased risk (Table IV.2). For a number of these occupations, namely loggers, electronic fitters, reinforced concreters, dockers, lorry and van drivers and labourers, risk estimates further increased when we restricted analyses to subjects employed for at least 10 years. With the exception of concreters and bricklayers, adjustment for educational level had limited impact on OR estimates. When we applied shrinkage through a semi-Bayesian approach, the occupational categories with the largest number of subjects remained associated with an increased OR, including painters, bricklayers, stonemasons and tile setters, bricklayers in construction industry, lorry and van drivers and labourers. Several occupations were associated with an OR of at least 2.0, including roofers (ISCO code: 953; OR: 2.04, 95% CI: 0.61-6.84, which decreased to 1.43 after SB adjustment), earth-moving and related machinery operators (ISCO: 974; OR: 2.12, 95% CI: 1.00-4.53, which decreased to 1.68 after SB adjustment), constructional steel erectors (ISCO: 87440; OR: 7.12, 95% CI: 0.86-59, which decreased to 1.56 after SB adjustment). Apart from lorry and van drivers, none of the other types of drivers had an increased UADT cancer risk (the OR for motor vehicle drivers as a whole, ISCO code 985, was 1.00, 95% CI: 0.77-1.30). The four 3-digit ISCO occupational categories associated with an increased risk of UADT cancer in the main analyses (Table IV.2) were analyzed also in the Paris centre. Risk of UADT cancer was increased among construction painters (ISCO code: 931; OR: 1.28, 95% CI: 0.54-3.02; 16 exposed cases) and bricklayers, stonemasons and tile setters (ISCO: 951, OR: 1.98, 95% CI: 1.01-3.87; 35 exposed cases), while the number of exposed subjects was too low to be analyzed for

loggers (four exposed cases and 0 exposed controls) and electronic fitters (two exposed cases and two exposed controls).

Out of 172 industries evaluated (four digits), we found an increased risk associated with a $p<0.05$ in nine categories (Table IV.2) and a decreased risk ($p<0.05$) in four categories. As summarized in Table IV.2, restriction of the analyses to subjects employed for at least 10 years increased most of the estimates, in particular for construction of motorways, roads, airfields and sport facilities (OR 5.61, 95% CI: 1.21-26.1; 15 exposed cases) and for cargo handling (OR: 4.85, 95% CI: 1.29-18.3; 14 exposed cases). Adjustment for educational level changed OR estimates more than marginally only for cattle and dairy farming industries. After semi-Bayesian shrinkage, cargo handling and categories related to the construction of buildings remained associated with an increased risk: Building of complete constructions or parts thereof, Civil engineering, not further specified, General construction of buildings and civil engineering works, Erection of roof covering and frames, Construction of motorways, roads, airfields and sport facilities. The SB adjustment decreased the OR associated with employment in the mining of uranium and thorium ores industry substantially from 9.4 to 1.5, as this category included only 10 subjects (nine of which came from the Prague centre). Among the industries associated with at least a two-fold increased risk, we found an OR of 2.56 (95% CI: 0.82-8.00) for workers in the manufacture of concrete products for construction purposes industry (NACE code: 2661), an OR of 2.33 (95% CI: 0.97-5.61) for manufacture of parts and accessories for motor vehicles and their engines (NACE: 3430) and an OR of 2.07 (95% CI: 0.58-7.34) for operation of dairies and cheese making (NACE: 1551). After SB adjustment all these OR estimates were below 1.5, with the exception of the manufacture of parts and accessories for motor

vehicles and their engines industry which was associated with an OR of 1.59 (95% CI: 0.82-3.08).

Table IV.3 reports the results for the increased-risk occupations and industries by UADT cancer subtype. For most of the occupations and industries there was no marked variation in risk estimates. However, bricklayers and workers employed in the farming of cattle and dairy farming industry had an increased risk only of oral/oropharyngeal and oesophageal cancer; the excess risks found for workers in the mining of uranium and thorium ores as well as for drivers were specific for cancer of the hypopharynx/larynx; the risk associated with having worked as a painter was higher for oral/oropharyngeal cancer; loggers had an increased risk especially for hypopharyngeal/laryngeal and oesophageal cancer.

We also carried out a full analysis in each of the three UADT cancer subtypes. This approach increases dramatically the number of comparisons but some results are of interest for the interpretation of excess risks found in the main analyses. For example, consistently with the results reported in Table IV.3, workers in the painting and glazing industry (NACE code: 4544) had an increased risk of oral/oropharyngeal cancer (OR: 2.03, 95% CI: 1.03-3.99; 23 exposed cases) but not of cancer in the hypopharyx/larynx (OR: 0.96, 95% CI: 0.41-2.25; 12 cases).

Table IV.2. Selected[a] occupations and industrial branches. Men

Category	Cases	Controls	Ever employment				At least 10 years of employment		Ever employment Semi-Bayes adjustment	
	N.	N.	OR1[b]	(95% CI)[b]	OR2[b]	(95% CI)[b]	OR1	(95% CI)	OR1[b]	(95% CI)
Occupation (ISCO code)[c]										
Loggers (631)	15	5	3.56	1.20-10.5	3.25	1.09-9.74	8vs.0	NA[b]	1.90	0.90-3.99
Electronics fitters (852)	15	9	2.45	1.01-5.95	2.55	1.04-6.22	3.23	0.73-14.3	1.73	0.89-3.37
Painters, construction (931)	52	23	1.74	1.01-3.00	1.68	0.97-2.90	1.49	0.73-3.03	1.58	0.98-2.54
Bricklayers, stonemasons and tile setters (951)	115	71	1.44	1.02-2.02	1.29	0.91-1.82	1.30	0.86-1.96	1.40	1.01-1.94
Electronics fitter -radio, television and radar equipment (85220)	10	4	3.93	1.14-13.6	4.16	1.19-14.5	3.41	0.56-20.6	1.82	0.83-4.00
Bricklayer, construction (95120)	73	50	1.58	1.05-2.37	1.42	0.94-2.13	1.48	0.91-2.42	1.50	1.03-2.19
Reinforced concreter (general) (95210)	17	4	3.46	1.11-10.8	2.94	0.94-9.19	5.62	0.69-46	1.82	0.85-3.87
Docker (97120)	16	7	2.91	1.05-8.05	2.94	1.07-8.06	4.11	0.88-19	1.77	0.87-3.64
Lorry and van driver (local transport) (98550)	112	68	1.46	1.04-2.05	1.36	0.97-1.90	2.12	1.24-3.62	1.42	1.03-1.96
Labourer (99910)	179	119	1.33	1.01-1.75	1.20	0.91-1.58	1.72	1.09-2.70	1.31	1.01-1.71
Industry (NACE code)[c]										
Farming of cattle, dairy farming (0121)	20	4	3.60	1.15-11.2	3.02	0.97-9.44	3.76	0.44-32	1.72	0.81-3.65
Mining of uranium and thorium ores (1200)	9	1	9.41	1.14-77.8	8.53	1.02-71.2	5vs.0	NA	1.48	0.60-3.67
Building of complete constructions or parts thereof; civil engineering - not further specified (452x)	73	45	1.62	1.05-2.48	1.45	0.94-2.23	1.25	0.75-2.07	1.50	1.01-1.22
General construction of buildings and civil engineering works (4521)	222	139	1.44	1.12-1.85	1.33	1.04-1.72	1.52	1.11-2.09	1.41	1.11-1.80
Erection of roof covering and frames (4522)	18	8	2.62	1.08-6.36	2.63	1.08-6.44	1.39	0.37-5.24	1.69	0.87-3.29
Construction of motorways, roads, airfields and sport facilities (4523)	23	7	3.03	1.23-7.46	2.84	1.14-7.06	5.61	1.21-26.1	1.82	0.93-3.56
Wholesale of wood, construction materials and sanitary equipment (5153)	9	5	3.33	1.00-11.1	3.39	1.02-11.2	3.34	0.49-23.0	1.61	0.75-3.48
Other retail sale in non-specialized stores (5212)	11	1	13.11	1.56-110	12.89	1.52-109	1vs.0		1.56	0.63-3.89
Cargo handling (6311)	26	10	2.60	1.17-5.75	2.47	1.12-5.47	4.85	1.29-18.30	1.77	0.95-3.31

[a] At least 10 exposed subjects, 95% confidence intervals excluding 1.
[b] OR1, odds ratio adjusted for center, age, alcohol and tobacco; OR2 adjusted as OR1 and education; CI, confidence intervals; NA, not applicable
[c] ISCO, International Standard Classification of Occupations; NACE, National Industrial Classification of All Economic Activities

Table IV.3. Selected[a] occupations and industrial branches by cancer site. Men

Category	Oral/Oropharynx (N=670)			Hypopharynx/Larynx (N=611)			Oesophagus (N=107)		
	N° of cases	OR[b]	95% CI[b]	N° of cases	OR	95% CI	N° of cases	OR	95% CI
Occupation (ISCO code)[c]									
Loggers (631)	3	1.75	0.36-8.44	8	5.01	1.51-16.6	3	7.06	1.24-40.3
Electronics fitters (852)	7	2.27	0.80-6.46	7	3.08	0.97-9.82	1	5.10	0.44-59.5
Painters, construction (931)	30	2.10	1.15-3.85	18	1.35	0.66-2.79	1	0.85	0.11-6.72
Bricklayers, stonemasons and tile setters (951)	58	1.71	1.14-2.56	43	0.96	0.60-1.53	12	2.14	1.03-4.46
electronics fitter -radio, television and radar equipment- (85220)	5	3.78	0.92-15.6	5	5.85	1.28-26.7	0	NA	
reinforced concreter (general) (95210)	7	3.81	1.05-13.8	8	4.35	1.16-16.3	0	NA	
Docker (97210)	7	2.12	0.63-7.15	6	2.40	0.67-8.57	1	14.3	1.34-154
lorry and van driver (local transport) (98550)	49	1.41	0.93-2.14	57	1.85	1.23-2.79	4	0.75	0.25-2.23
Labourer (99910)	83	1.26	0.90-1.76	72	1.37	0.96-1.95	14	1.64	0.84-3.20
Bricklayer –construction- (95120)	38	1.78	1.11-2.87	22	0.99	0.56-1.77	11	3.03	1.37-6.72
Industry (NACE code)[c]									
Farming of cattle, dairy farming (0121)	10	3.45	0.97-12.3	4	1.03	0.21-5.04	4	6.88	1.44-32.8
Mining of uranium and thorium ores (1200)	0	NA[b]		7	21.5	2.47-188	1vs.0	NA	
Building of complete constructions or parts thereof; civil engineering - not further specified (452x)	28	1.81	1.05-3.11	33	1.64	0.97-2.79	12	2.17	1.03-4.57
General construction of buildings and civil engineering works (4521)	108	1.56	1.16-2.11	91	1.30	0.94-1.80	11	0.90	0.44-1.82
Erection of roof covering and frames (4522)	6	1.84	0.60-5.66	8	2.76	0.96-7.99	1	3.72	0.40-34.2
Construction of motorways, roads, airfields and sport facilities (4523)	16	5.15	1.97-13.5	6	2.10	0.61-7.24	1	1.67	0.1716.4
Wholesale of wood, construction materials and sanitary equipment (5153)	3	2.21	0.42-11.6	4	4.81	1.17-19.7	2	20.0	2.17-185
Other retail sale in non-specialized stores (5212)	6	14.2	1.4-145	4	10.8	110	1vs.0	NA	
Cargo handling (6311)	9	1.68	0.63-4.49	13	2.60	1.00-6.75	2	7.52	1.29

[a] Occupations and industrial branches selected for Table IV.2
[b] OR, odds ratio adjusted for center, age, alcohol and tobacco; CI, confidence intervals; NA, not applicable
[c] ISCO, International Standard Classification of Occupations; NACE, National Industrial Classification of All Economic Activities

Women

In total, 71 occupations (3 and 5 ISCO digits) and 44 industries (4 NACE digits) had at least 10 exposed subjects and were therefore retained for further analyses. Among these, employment in the retail sale of furniture, lighting equipment and household articles not elsewhere classified (NACE code: 5244) was the only category associated with an increased risk with a p-value<0.05 (OR: 3.53, 95% CI: 1.24-10.07; 12 exposed cases). Out of the 19 high-risk occupations and industries found in the analyses restricted to men, having worked as a labourer (ISCO code: 99910; tobacco- and alcohol-adjusted OR: 1.44; 95% CI: 0.68-3.03; 18 exposed cases), having been employed in the general construction of buildings and civil engineering works industry (NACE code: 4521, OR: 0.52; 95% CI: 0.16-1.73; 4 exposed cases) and having worked in the other retail sale in non-specialized stores industry (NACE code: 5212, OR: 1.97, 95% CI: 0.88-4.40; 16 exposed cases) included a sufficient number of subjects (at least 10) to be investigated.

Discussion

We studied occupational history in relation to UADT cancer risk in a large multicentre European study. Information on occupational history was obtained in a face-to-face interview using a standardized and detailed questionnaire, which was then coded by trained coders blinded to the case-control status of the subjects. We also had detailed information on the main potential confounders, namely tobacco and alcohol, which have been analysed in ad-hoc papers[164].

The assessment of exposure was based on standard coding of occupations and industries based on ISCO-NACE classification systems, implying that no direct information on specific carcinogens was available. Being an exploratory study assessing a large number

of potential associations, there is the risk of false positive associations, as well as the possibility that some of the occupations and industries which were not associated with UADT cancer risk still entail exposure to UADT carcinogens. To evaluate the robustness of our positive findings, we used SB adjustment and conducted analyses restricted to occupations and industries in which subjects had worked for at least 10 years. Nevertheless, results should be interpreted with caution and discussed in the context of previous knowledge on occupational risk factors for UADT cancer. Selection and recall bias are other potential limitations of our study. Participation in the study was lower among controls than cases especially in the centres which used a population-based design. However, adjustment for educational level, typically one of the main determinants of participation, modified only marginally our estimates suggesting a limited role of selection bias. Although we cannot exclude the possibility of recall bias, occupational factors are not established causes of UADT cancer and it is therefore unlikely that cases reported their occupational history more accurately or in a biased fashion. Consistently, cases and controls reported a similar number of job periods (Table IV.1).

Some of our findings are consistent with previous studies on occupational factors for UADT cancer; these results should therefore be considered as supportive of previously reported associations, although in most of the cases the involved carcinogens are not known. We found an excess risk of UADT cancer for some categories of construction workers, including reinforced concreters, bricklayers, constructional steel erectors, roofers and workers employed in the erection of roofs and frames, and those working in the construction of roads. Some of these risks increased in analyses restricted to long duration of employment and many of them remained elevated after SB adjustment.

Bricklayers had an increased risk of oral/oropharyngeal and oesophageal cancer but not of the laryngeal cancer. This is consistent with recent results from a comprehensive register-based study carried out in Nordic countries[161] that found a 30–40% increased risk for cancer of the oral cavity (87 exposed cases) and the pharynx (115 cases), while relative risks of 1.11 for oesophageal cancer (180 cases) and 1.05 for laryngeal cancer (167 cases). Two previous studies on laryngeal cancer that investigated bricklayers separately (with about 50 exposed cases each) found relative risks of 1.03[150] and 1.6[149].

The IARC international cohort on asphalt workers[167] and a recent update limited to the German part of that cohort[168] reported an increased risk of UADT cancer which is consistent with our findings for workers in the erection of roofs and construction of roads, although results from the cohort study were not adjusted for smoking and alcohol and the increased risk was mainly due to the German data. There is little information on roofers and pavers from previous case-control studies, with the exceptions of two studies on laryngeal cancer reporting a relative risk of 0.4, based on five exposed cases[149], and a relative risk of 6.4 based on 22 exposed cases[147].

The evidence on the risk of UADT cancer among concrete workers is more convincing. In particular, our finding of an increased risk for reinforced concreters is consistent with similar findings from two previous case-control studies on laryngeal cancer[149, 150] and an increased risk in the concrete and cement manufacture industry found in another case-control study on laryngeal cancer[156]. Some studies report an association between exposure to cement dust, which is a complex and heterogeneous mixture, and

cancer of the larynx[169] or pharynx[170], although other studies did not replicate this association[156, 163].

Apart from specific exposure to cement dust in concrete workers, employment in the different types of construction jobs that we found at increased risk of UADT cancer involves exposure to a number of agents which have been previously reported to be associated with at least one of the different UADT subtypes, including asbestos, polycyclic aromatic hydrocarbons, inorganic dusts and solvents[162, 163].

The excess risk for painters found in our study remained after SB-adjustment while it did not increase with increasing duration of employment. The risk was evident mainly for oral/oropharyngeal cancer. Several studies have investigated UADT cancer risk, especially laryngeal cancer, in association with having worked as a painter, finding moderately increased risks[145-150, 155, 157, 158, 160]. Similarly, some cohort studies found increased risk of a small magnitude for cancers of the pharynx or oral cavity[161, 171-173]. A recent IARC monograph has summarized the epidemiological evidence for cancer risk in painters, including the risk of UADT cancer[172]. The working group concluded that, although data were insufficient for evaluation, there was some consistency between case-control and cohort studies for an increased risk of cancers of the pharynx and oesophagus. Painters are or were exposed to a great number of chemical compounds, including organic and inorganic solvents, chromium, pigments, additives, binders as well as silica and asbestos and it is thus difficult to speculate on possible specific UADT carcinogens which could explain the association[172].

The risk of laryngeal cancer that we found among uranium miners was entirely due to one of the participating centres, namely the city of Prague that has uranium mines in its vicinity. The collaborative analysis of cohort studies of underground miners exposed to radon published in 1995[174] and some more recent analyses of cohorts of uranium miners[175, 176] found a slightly increased risk of laryngeal cancer of about 20%[177].

Drivers have been found to have an increased risk of UADT cancer in a number of studies although there is marked inconsistency[144, 148, 150, 151, 157, 159, 160]. In our study the risk was increased for lorry and van drivers as well as for earth-moving and related machinery operators but not for drivers as a whole. The increased risk among loggers observed in the present study has been found before in a large case-control study on laryngeal cancer[150]. However, in the recent register-based study in the Nordic countries, the incidence of each UADT cancer type was decreased among forestry workers compared with the general population[161]. Labourers, as well as dockers, have been noted to have an increased UADT cancer risk in a number of studies but these categories are rather heterogeneous[145, 147, 149-152, 154, 155].

The increased risk that we found among workers employed in the cattle and dairy farming industries is a new association and should then be treated with a greater degree of caution. The association was slightly attenuated after adjustment for educational level. An increased risk of UADT cancer was also found among workers employed in the operation of dairies and cheese making industry (NACE: 1551, OR: 2.07) and for dairy farm workers (ISCO 625, OR: 1.75) and dairy product processors (ISCO 775, OR: 1.83). Most, if not all, previous studies on UADT cancer did not analyse dairy workers separately. In a Finnish cohort study on cancer risk among food industry workers, the

risk was increased by 30% for laryngeal cancer (three cases) and by 100% for oesophageal cancer (three cases)[178]. Cohort studies of farmers including a large proportion of dairy farmers do not reveal an increased risk of UADT cancer[179-181].

There is little information on occupational risk factors for UADT cancer in women. Our study included almost 400 female cases and a previous case-control study on UADT cancer[158] included 350 women but prevalences of exposure were low and neither study found a clear evidence of an increased risk associated with specific occupations or industries.

In conclusion, this large European study provides evidence that occupational exposures play a role in UADT cancer aetiology and contribute to explain, together with alcohol, smoking and diet, socioeconomic differences typically observed in UADT cancer risk[142]. The most internally consistent findings, also supported by previous studies, were obtained for some specific workers employed in the construction industry, including reinforced concreters, bricklayers, painters and workers employed in the construction of roads or the erection of roofs.

Acknowledgements

The Dublin centre acknowledges the clinical support of Prof. J Reynolds, Prof. C. Timon and their colleagues. The Aviano centre thanks Mrs. O. Volpato for study coordination, Ms. G. Bessega, L. Zaina, for their help in data collection and Drs. S. Sulfaro and D. Politi for providing hospital case patients. The Manchester centre is grateful to the support of many clinicians and staff of the hospitals, interviewers, data managers, pathology departments, and primary care clinics and acknowledges the help

of Dr. Ann-Marie Biggs and Professor Martin Tickle in study conducted in the Manchester centre and Professor Phil Sloan and Professor Nalin Thakker who in addition co-ordinated sample collection and processing for all the UK centres. G.J.M. and T.V.M. partly worked on this study while at the University of Manchester.

CHAPTER V

Hierarchical regression for multiple comparisons in a case-control study of occupational risks for lung cancer

Marine Corbin, Lorenzo Richiardi, Roel Vermeulen, Hans Kromhout, Franco Merletti, Susan Peters, Lorenzo Simonato, Kyle Steenland, Neil Pearce and Milena Maule

Background: Occupational studies often involve multiple comparisons and therefore suffer from false positive findings. Semi-Bayes adjustment methods have sometimes been used to address this issue. Hierarchical regression is a more general approach, including Semi-Bayes adjustment as a special case, that aims at improving the validity of standard maximum-likelihood estimates in the presence of multiple comparisons by incorporating similarities between the exposures of interest in a second-stage model.

Methodology/Principal Findings: We re-analysed data from an occupational case-control study of lung cancer, applying hierarchical regression. In the second-stage model, we included the exposure to three known lung carcinogens (asbestos, chromium and silica) for each occupation, under the assumption that occupations entailing similar carcinogenic exposures are associated with similar risks of lung cancer. Hierarchical regression estimates had smaller confidence intervals than maximum-likelihood estimates. The shrinkage toward the null was stronger for extreme, less stable estimates (e.g., ''specialised farmers'': maximum-likelihood OR: 3.44, 95%CI 0.90-13.17; hierarchical regression OR: 1.53, 95%CI 0.63-3.68). Unlike Semi-Bayes adjustment toward the global mean, hierarchical regression did not shrink all the ORs towards the null (e.g., ''Metal smelting, converting and refining furnacemen'': maximum-likelihood OR: 1.07, Semi-Bayes OR: 1.06, hierarchical regression OR: 1.26).

Conclusions/Significance: Hierarchical regression could be a valuable tool in occupational studies in which disease risk is estimated for a large amount of occupations when we have information available on the key carcinogenic exposures involved in each occupation. With the constant progress in exposure

assessment methods in occupational settings and the availability of Job Exposure Matrices, it should become easier to apply this approach.

Chapter based on PLoS ONE. 2012; 7(6): e38944. doi:10.1371/journal.pone.0038944

Introduction

Occupational studies often involve the simultaneous analysis of multiple exposures and/or multiple occupations. A conventional approach to such analyses is to build a separate model for each occupation, adjusting for possible confounders. However, this approach treats all the associations equally, without accounting for the fact that some occupations are *a priori* more likely to be at risk than others, i.e. that some occupations have prior evidence of associations with the disease under study, whereas other occupations do not. Furthermore, for those occupations which show strongly elevated (or reduced) relative risks, their risk estimates may be biased away from the null due to random error, and it is likely that if the study were repeated, then risk estimates closer to the null would be found, due to 'regression to the mean'.

Semi-Bayes adjustment methods have been shown to be valid approaches to these problems, particularly when the parameters to be estimated can be categorised into groups within which the various occupations or exposures have risks which are similar or ''exchangeable'' on the basis of *a priori* knowledge[182]. The basic idea of Semi-Bayes adjustment for multiple comparisons is that the observed variation of the estimated risks around their geometric mean will be larger than the variation of the true (but unknown) risks. The Semi-Bayes method[48] specifies an *a priori* value for the variation of the true risks; this *a priori* value is then used to adjust the observed risks[43]. The adjustment consists in shrinking outlying estimates towards the overall

mean of the observed estimates. The larger the individual variance of the estimates, the stronger is the shrinkage, i.e. the shrinkage is stronger for less reliable estimates based on small numbers.

Semi-Bayes adjustment is a special case of the more general method of hierarchical regression[37]. The latter approach incorporates a number of specific types of regression model as special cases including Bayesian regression, Semi-Bayes regression, Stein regression, penalized likelihood regression, and ridge regression. In the current context, hierarchical regression can be used to incorporate prior similarities between the exposures of interest in a second-stage model. This approach has been used previously in several studies involving the assessment of multiple exposures/risk factors, e.g. studies on diet[53], genetic studies[49, 50] and occupational studies[51, 58, 183]. The objective of the present work was to re-analyse data from an occupational case-control study of lung cancer, applying hierarchical regression and including prior information from a validated Job-Exposure-Matrix (JEM). In particular, we included in the second-stage model the exposure to three known lung carcinogens for each occupation, under the assumption that occupations entailing similar exposure levels to the same lung carcinogen are associated with similar risks of lung cancer.

Materials and Methods

Ethics Statement

The current study is a re-analysis of the Italian subset of the multicentric study on lung cancer from the International Agency for Research on Cancer (IARC)[44], hence no additional ethical committee approval was requested.

Description of the Data

The data are from a population-based case-control study conducted between 1990 and 1992 in two areas of Italy (the city of Turin and the Eastern part of Veneto Region). The study methodology has been described elsewhere[44]. Briefly, cases (956 men and 176 women) were all individuals diagnosed with incident primary lung cancer during 1990-1992, aged less than 75 and resident in the study areas. Controls (1,253 men and 300 women) were randomly selected from the local population registries and frequency matched with cases by gender, study area and five-year age groups. Information was collected on basic demographic details, active and passive smoking, and lifetime occupational history. In particular, the dates of beginning and ending work, as well as the job title and branch of industry, were recorded for each occupational period that lasted at least 6 months. Job titles and branches of industry were coded blind to case-control status according to the International Standard Classification of Occupations (ISCO-68)[45] and the International Standard Industrial Classification (ISIC)[46], respectively. The current analyses were carried out only in men.

We focused on three chemicals which were classified by the International Agency for Research on Cancer (IARC)[184] as group 1 carcinogens targeting the lung: asbestos, chromium and silica. Exposure to these carcinogens was assessed through a General Population Job Exposure Matrix (DOM-JEM) developed in 2010 by three occupational experts (HK, RV and SP) for a large pooled case-control study on lung cancer[185]. The DOM-JEM assigns an ordinal exposure score for several lung carcinogens (0 =no exposure, 1= low exposure, 2= high exposure) to each ISCO code.

Conventional Analysis

The analyses were done at the three-digit ISCO code level. For the ISCO codes starting by ‘‘X’’ (workers not classifiable by occupation) and for those specified to a maximum of 2 digits, all the corresponding occupational histories were deleted from the dataset, resulting in the exclusion of 5 cases and 14 controls. Only job-codes with at least ten subjects were retained in the analyses (n=129). The first-stage models estimated the risk of lung cancer for each of the 129 occupations separately. The Odds Ratio (OR) for ever being exposed to each job was modelled using unconditional logistic regression, adjusting for age, study area and cigarette smoking status (never, ex, current):

$$\text{logit}\left[P\left(Y=1|\text{occ}_\text{i}, w\right)\right]=\alpha_i+\text{occ}_\text{i}\beta_i+\mathbf{w}\gamma_i \qquad (1)$$

where Y is a dichotomous variable representing the lung cancer status (Y=1: cases; Y=0: controls), occ_i (i = 1,…,129) is a dichotomous variable representing the exposure status to the i^{th} occupation, **w** is a vector of covariates included in the model (i.e. age, study area, and cigarette smoking status), α_i is the intercept term, β_i is the regression coefficient corresponding to the i^{th} occupation, and γ_i is the vector of regression coefficients corresponding to the covariates for the i^{th} occupation.

We also carried out conditional logistic regression. Since the estimates obtained through conditional and unconditional regression adjusting for matching variables were very similar, here we show only those obtained through unconditional logistic regression.

The ORs with corresponding 95% confidence intervals (CI) were estimated through maximum-likelihood using the SAS Logistic procedure.

Hierarchical Regression

Hierarchical regression can be used to attempt to improve on maximum-likelihood (ML) estimates by using a second-stage linear model[49, 53]. The second-stage model used here regressed the ln(OR)s of the occupations on the occupations' estimated exposure levels to asbestos, chromium and silica.

$$\beta = \mathbf{Z}\pi + U$$

$$U \sim \mathrm{N}\left(\mathbf{0}, \tau^2 \mathbf{T}\right) \qquad (2)$$

β is the 129-element vector of the ln(OR)s for the occupations.

$\mathbf{Z}$ is the 129×7 matrix (intercept and 2 indicator variables per exposure) obtained from the DOM-JEM[185] that classifies the 129 occupations according to their levels of exposure to asbestos, chromium and silica. Each carcinogen has two possible levels of exposure, expressed by two dichotomous variables.

More specifically, we have:

$\mathbf{Z}_{i0} = 1$ are the elements of the 1^{st} column and the intercepts of the model,

$\mathbf{Z}_{ij_k} = \begin{cases} 1 \text{ if the } i^{\text{th}} \text{ occupation entails a level} = k \text{ exposure to the } j^{\text{th}} \text{ carcinogen} \\ 0 \text{ otherwise} \end{cases}$ is the value at the

i^{th} row and $\left(j \times 2 + k - 1\right)^{\text{th}}$ column, where $j \in \{1,2,3\}$, $k \in \{1,2\}$, and $\mathbf{Z}_{ij_1}$ and $\mathbf{Z}_{ij_2}$ are mutually exclusive.

Appendix AII.1 shows rows 55 to 60 of matrix $\mathbf{Z}$. For example, $\mathbf{Z}_{55,3_1}$ is located at the 55^{th} row and the 6^{th} column of the matrix and equals 1 because "nursery workers and gardeners" are exposed to silica (from soil) at level 1.

π is the 7-element vector (estimated by the second-stage model) of the coefficients corresponding to the effects on lung cancer of the levels of exposures to the three carcinogens described in $\mathbf{Z}$.

U is a 129-element vector of the error terms representing the residual effect of being employed in each occupation after accounting for the exposure to asbestos, chromium and silica.

$\mathbf{0}$ is a 129-element vector of zeros.

$\tau^2\mathbf{T}$ is the 129×129 second-stage covariance matrix. The second-stage variance for an estimate for a particular occupation represents the residual variance of the effect of the occupation after taking into account the effects of the three lung carcinogens. This can be estimated from the data (Empirical Bayes) or specified *a priori* (Semi-Bayes). We used here the Semi-Bayes approach. τ is a parameter used to control the strength of the common shrinkage of all the ML coefficients towards their prior means. We set τ to 0.23, 0.41, 0.59 and 0.76, corresponding to the assumptions that 95% of relative risks would lie within a 2.5, 5, 10 and 20 fold-range of each other, respectively, if $\mathbf{T}$ was the identity matrix. We assumed that the second-stage variance for each occupation depends on its levels of exposure to the three carcinogens, so that the higher the levels of exposure, the smaller the second-stage variance. For ease of computation, $\tau^2\mathbf{T}$ did not include residual correlation between occupations. $\mathbf{T}$ is then a diagonal matrix (see Appendix AII.1 for examples of calculation) with:

$$t_{ii} = \exp\left(-\frac{1}{3}\sum_{j=1}^{3}\sum_{k=1}^{2} k\mathbf{Z}_{ij_k}\right) \qquad (3)$$

The model was fitted with R (free software for statistical computing and graphics)[90] (although such analyses can also be done in SAS and Stata, or with any logistic regression package by adding simple prior data[186]). The code is a modified version of the code provided by Chen and Witte[49] and is available in Appendix III. The coefficients π were estimated through weighted least squares (see Chapter II and Appendix II). Substituting them back into the equation (2) yielded prior means $\mathbf{Z}\tilde{\pi}$ for the occupations' coefficients β_i. Hierarchical regression estimates (posterior estimates) for the coefficients for each occupation were then obtained by averaging the ML coefficients (from conventional analysis) and their respective prior means, so that the larger the diagonal elements of $\tau^2\mathbf{T}$, the stronger the shrinkage of coefficients towards their prior means.

Semi-Bayes Adjustment towards the Global Mean

We compared the results obtained through hierarchical regression with those obtained through a more traditional Semi- Bayes adjustment towards the global mean, used previously in occupational studies involving multiple comparisons[3, 4, 17, 43, 60, 61]. The variance of the true ln(OR)s was assumed to be 0.25. Assuming a normal distribution of the ln(OR)s, this choice implies that the true ORs are within a 7-fold range of each other[48]. The Semi-Bayes adjustment was applied separately within two groups of occupations believed to entail different levels of exposure to lung carcinogens: the occupations held by white-collar workers (identified by the first digit of ISCO code<6, less likely to entail exposure to carcinogens) and the occupations held by blue-collar workers (identified by the first digit of ISCO code≥6, more likely to entail some or heavy exposure to carcinogens). For each group of occupations, this

method was equivalent to a particular case of hierarchical regression in which only the intercept was included in the second-stage model.

Results

Table V.1 summarizes the basic characteristics of the subjects included in our analyses.

Table V.1. Selected characteristics of cases and controls

	Cases		Controls	
	N	(%)	N	(%)
Center				
Turin	482	(50.7)	669	(54.1)
Eastern Veneto region	469	(49.3)	568	(45.9)
Age, years				
Mean, (Standard Deviation)	62.3	(7.4)	63.3	(7.7)
Cigarette smoking				
Never smoker	15	(1.6)	248	(20.0)
Ex-smoker	327	(34.4)	587	(47.5)
Current smoker	609	(64.0)	402	(32.5)
Total	951		1237	

Table V.2 presents the ORs of lung cancer for ever being exposed to each level of exposure of the carcinogens included in the second-stage model (asbestos, chromium and silica). These ORs were estimated through logistic regression models, adjusting for age, study area and cigarette smoking status (never, ex, current). Ever being exposed to each of the three carcinogens was associated with an increased risk of lung cancer, with higher risks observed for high levels of exposure.

Table V.2. Odds ratio (OR) of lung cancer and 95% confidence intervals (CI) for ever being exposed to each level of exposure of asbestos, chromium and silica

Carcinogen	Exposure level	Cases/Controls	OR[95%CI][a]
Asbestos	**Unexposed (0)**	429/682	1.00
	Ever low (1)	477/512	1.43[1.18-1.73]
	Ever high (2)	45/43	1.62[1.01-2.61]
Chromium	**Unexposed (0)**	579/808	1.00
	Ever low (1)	270/339	1.11[0.90-1.37]
	Ever high (2)	102/90	1.55[1.11-2.15]
Silica	**Unexposed (0)**	627/862	1.00
	Ever low (1)	288/345	1.19[0.97-1.46]
	Ever high (2)	36/30	1.58[0.92-2.71]

[a] Estimated through logistic regression models, adjusting for age, study area and cigarette smoking status (never, ex, current).

Table V.3 shows the descriptive statistics for the distribution of the 129 ln(OR)s obtained through ML estimation, Semi-Bayes (SB) adjustment, and Hierarchical Regression (HR) with τ =0.76, τ =0.59, τ =0.41 and τ =0.23.

Compared with ML, the mean of the distribution of the ln(OR)s is pulled towards zero after SB and HR. For HR, this effect is stronger for smaller values of τ. The standard deviation of the distribution of the ln(OR)s is also reduced by both SB and HR and is smaller for smaller values of τ (Table V.3). It can also be noted that both SB and HR estimates have on average smaller standard errors.

The kernel density plots (Figure V.1) of the ln(OR)s show less left skewed distributions for SB and HR than for the ML estimates (smaller medians after SB and HR are also apparent in Table V.3). This is due to the fact that the extreme estimates, which are more likely to be unstable, are pulled towards their prior means.

In Table V.3, we can see that, for SB, the mean and the standard deviation of the ln(OR)s distribution are included between the corresponding values for HR[τ =0.41]

and HR[τ =0.59]. However, the distribution obtained after SB is more left skewed than after HR (Figure V.1). The density curve for SB has a higher slope on its right side than on its left side: while the left side lies between the curves for HR[τ =0.41] and HR[τ =0.59], the right side lies under both curves. This indicates that extreme positive estimates are in general shrunk more strongly towards the null value (ln(OR)=0) through SB than through HR.

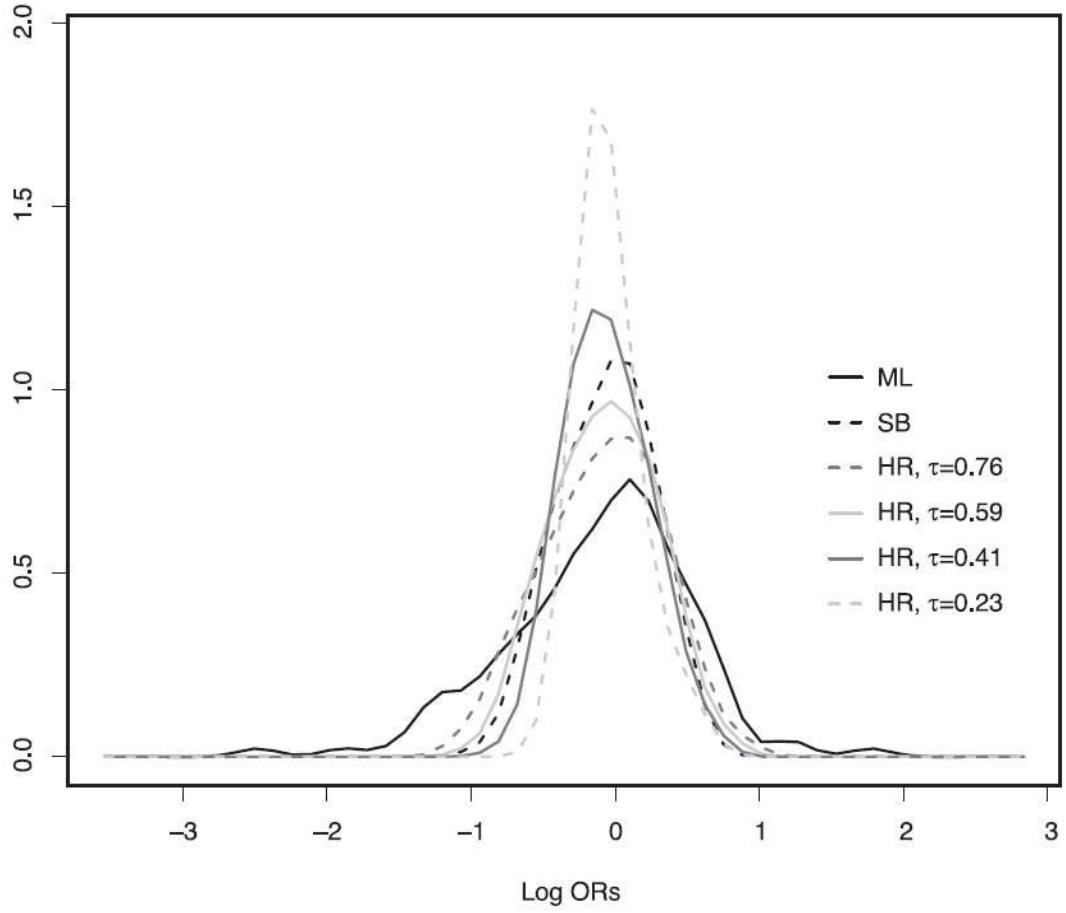
Table V.3. Descriptive statistics for the distribution of the ln(OR)s of lung cancer for the 129 selected occupations (3-digit ISCO codes; n>10) obtained using Maximum Likelihood (ML), Semi-Bayes adjustment towards the global mean (SB) and hierarchical regression (HR)

	ML	SB	HR			
			τ = 0.76	τ = 0.59	τ = 0.41	τ = 0.23
Maximum prior range of ORs' variation		7-fold range[a]	20-fold range[b]	10-fold range[b]	5-fold range[b]	2.5-fold range[b]
Mean of estimated ln(OR)s	-0.12	-0.07	-0.08	-0.07	-0.06	-0.04
Median of estimated ln(OR)s	-0.03	-0.06	-0.06	-0.07	-0.08	-0.08
Standard deviation of estimated ln(OR)s	0.63	0.31	0.41	0.35	0.28	0.20
Mean of estimated standard errors	0.45	0.32	0.37	0.34	0.28	0.20

[a] A 7-fold range means that we assume *a priori* that 95% of the true ORs are within a 7-fold range of each other (e.g. from 0.5 to 3.5).

[b] Prior range of ORs' variation when matrix $\mathbf{T}$ is the Identity matrix. A 20, 10, 5, or 2.5-fold range means that we have 95% *a priori* certainty that the residual OR for being ever employed in an occupation after accounting for the effect of the carcinogens will lie within a 20, 10, 5 or 2,5-fold range.

Figure V.1. Kernel density distributions of the ln(OR)s. Kernel density distributions of the ln(OR)s of lung cancer for the 129 selected occupations obtained using Maximum Likelihood (ML), Semi-Bayes adjustment towards the global mean (SB) and hierarchical regression (HR)



The effect of the shrinkage can be seen in the scatter plots in Figure V.2, where the ORs for each occupation estimated with HR and SB are plotted against ML estimates. The further ML estimates are from the null value (OR=1), the more scattered are HR and SB estimates and the stronger is the shrinkage. As expected, extreme estimates are pulled more strongly for smaller values of τ .

Table V.4 reports the OR estimates obtained through the different methods for the occupations associated with the twenty highest risks of lung cancer in the conventional analysis. Shrinkage is particularly strong for specialised farmers (ML OR=3.44, SB OR=1.59, HR OR[τ =0.76]=1.81, HR OR[τ =0.23]=1.00) and for ships' engine-room ratings, who are highly exposed to asbestos (ML OR=5.88, SB OR=1.54, HR OR[τ =0.76]=2.43, HR OR[τ =0.23]=1.78). This is due to the fact that these two occupations are held by a small number of subjects and the confidence intervals for the ML estimates are therefore very large. Despite the large CIs, however, the 'shrunk' estimates still indicate that these occupations are associated with an increased risk of lung cancer, and their ORs are consistent with those of other occupations which involve exposure to lung carcinogens.

SB with an *a priori* true standard deviation of 0.5 provided estimates that were less scattered than the HR estimates obtained with the chosen values of τ (Figure V.2). In particular, SB shrunk all the increased ML estimates towards the null, whereas some increased estimates were pulled away from the null when using HR. For example, the ML risk estimate for "Metal smelting, converting and refining furnacemen" (ML OR=1.07, SB OR=1.06, HR OR[τ =0.59]=1.26, HR OR[τ =0.41]=1.37) is close to the null whilst HR, weighting for their exposure to both asbestos (low exposure) and chromium (high exposure), pulls the risk estimate away from the null. Similarly, HR estimates a higher risk for "Miners and quarrymen" (ML OR=1.19, SB OR=1.14, HR OR[τ =0.59]=1.27, HR OR[τ =0.41]=1.30), exposed to both asbestos (low exposure) and silica (high exposure). "Metal annealers, temperers and case-hardeners" (ML OR=1.14, SB OR=1.08, HR OR[τ =0.59]=1.42, HR OR[τ =0.41]=1.44) are only highly exposed to chromium and "Railway engine drivers and firemen" (ML OR=0.97, SB

OR=1.01, HR OR[τ =0.59]=1.35, HR OR[τ =0.41]=1.47) are only highly exposed to asbestos. However, the ML estimates have large variances, which increases the strength of the shrinkage towards the prior ORs and results in elevated risk estimates after HR. On the other side SB, using less informative priors, performs a more systematic shrinkage and results in a general reduction of the ORs. Some ML ORs below 1 are also shrunk above 1 by HR whereas they are shrunk upwards but below 1 by SB, as in the case of "Metal casters" (ML OR=0.58, SB OR=0.84, HR OR[τ =0.76]=0.91, HR OR[τ =0.59]=0.98, HR OR[τ =0.41]=1.07, HR OR[τ =0.23]=1.12). Therefore, in general, SB with an *a priori* true standard deviation of 0.5 and HR with τ =0.59 provide shrinkages of similar magnitude, but different risk estimates for occupations known *a priori* to be exposed to lung carcinogens.

Figure V.2. Relationship between the ORs obtained with the different approaches. Scatter plots of the ORs of lung cancer for the 129 selected occupations estimated using hierarchical regression (HR) with τ = 0.76 vs. Maximum Likelihood (ML) (A), HR with τ = 0.59 *vs*. ML (B), HR with τ = 0.23 vs. ML (C) and Semi-Bayes adjustment towards the global mean (SB) *vs.* ML (D)

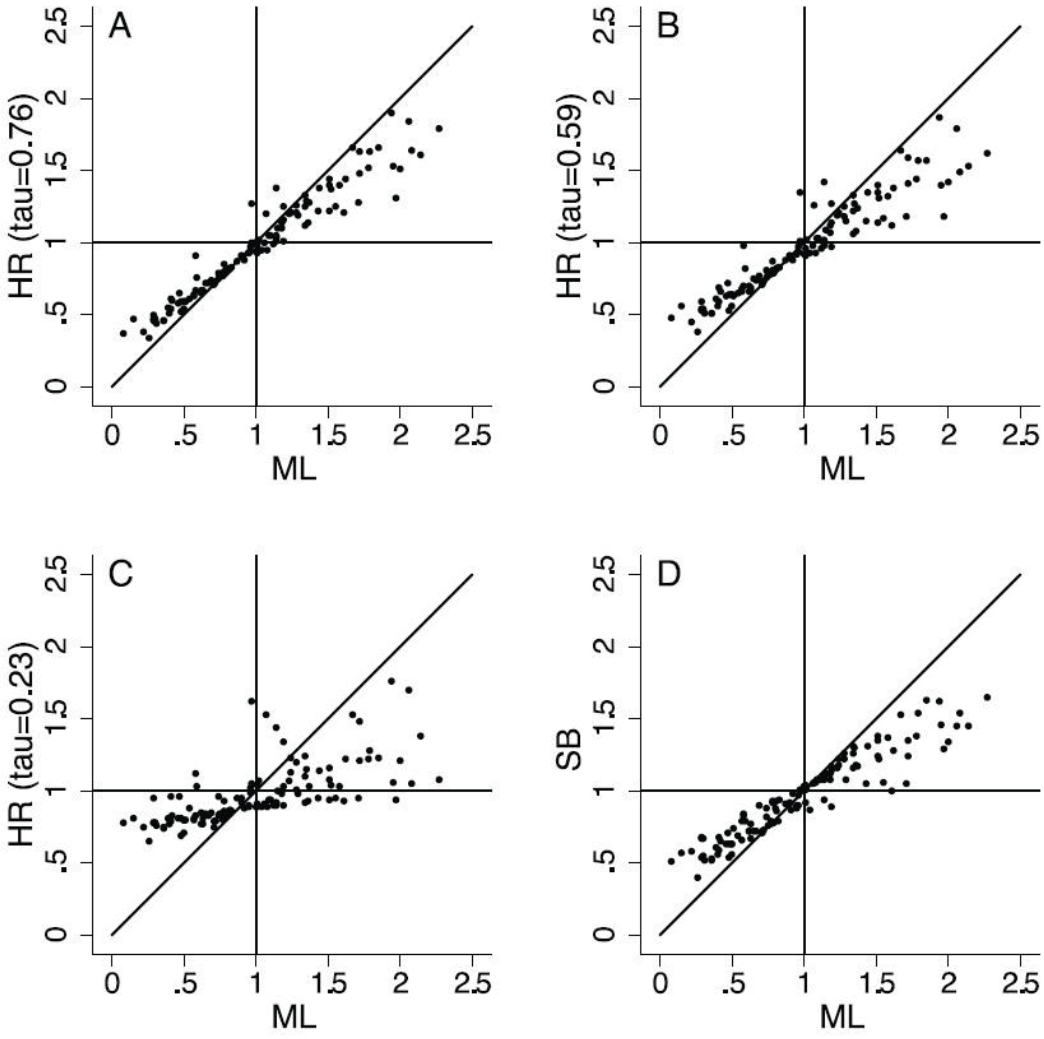


Table V.4. ORs of lung cancer and 95% confidence intervals obtained using Maximum Likelihood (ML), Semi-Bayes adjustment towards the global mean (SB) and hierarchical regression (HR) for the occupations associated with the twenty highest ORs in the conventional ML analysis

ISCO CODE - OCCUPATION	CASES / CONTROLS	ML	SB	HR				CARCINOGENIC EXPOSURE		
			7-fold range	τ = 0.76 *(20-fold range)*	τ = 0.59 *(10-fold range)*	τ = 0.41 *(5-fold range)*	τ = 0.23 *(2.5-fold range)*			
		OR[95%CI]	OR[95%CI]	OR[95%CI]	OR[95%CI]	OR[95%CI]	OR[95%CI]	ASB[a]	CR[a]	SI[a]
034-Electrical and electronics engineering technicians	5/9	1.61[0.44-5.88]	1.00[0.45-2.21]	1.21[0.45-3.22]	1.12[0.47-2.66]	1.02[0.51-2.02]	0.93[0.60-1.44]	0	0	0
628-Farm machinery operators	9/8	1.62[0.55-4.81]	1.28[0.62-2.65]	1.44[0.58-3.54]	1.38[0.59-3.22]	1.31[0.61-2.81]	1.22[0.62-2.40]	0	0	2
872-Welders and flame-cutters	47/37	1.67[1.03-2.71]	1.53[0.99-2.36]	1.66[1.06-2.59]	1.64[1.06-2.53]	1.60[1.07-2.41]	1.53[1.06-2.19]	0	2	0
039-Engineering technicians not elsewhere classified	7/7	1.71[0.51-5.72]	1.05[0.49-2.28]	1.28[0.50-3.28]	1.18[0.51-2.72]	1.05[0.54-2.07]	0.95[0.62-1.46]	0	0	0
727-Metal drawers and extruders	5/5	1.72[0.45-6.63]	1.24[0.56-2.76]	1.63[0.66-4.03]	1.59[0.72-3.52]	1.54[0.81-2.91]	1.48[0.94-2.33]	0	2	0
725-Metal moulders and coremakers	10/9	1.72[0.65-4.58]	1.35[0.67-2.70]	1.48[0.70-3.13]	1.41[0.72-2.76]	1.32[0.77-2.28]	1.21[0.83-1.77]	0	1	1
952-Reinforced-concreters, cement finishers and terrazzo workers	13/8	1.78[0.70-4.56]	1.38[0.70-2.74]	1.52[0.73-3.16]	1.44[0.75-2.78]	1.34[0.78-2.30]	1.22[0.84-1.77]	0	1	1
729-Metal processers not elsewhere classified	28/20	1.79[0.97-3.32]	1.54[0.91-2.61]	1.63[0.95-2.82]	1.57[0.94-2.62]	1.45[0.92-2.28]	1.28[0.91-1.80]	0	1	1

ISCO CODE - OCCUPATION	CASES / CONTROLS	ML	SB	HR				CARCINOGENIC EXPOSURE		
			7-fold range	**τ = 0.76** *(20-fold range)*	**τ = 0.59** *(10-fold range)*	**τ = 0.41** *(5-fold range)*	**τ = 0.23** *(2.5-fold range)*			
		OR[95%CI]	**OR[95%CI]**	**OR[95%CI]**	**OR[95%CI]**	**OR[95%CI]**	**OR[95%CI]**	**ASB[a]**	**CR[a]**	**SI[a]**
931-Painters, construction	42/29	1.85[1.09-3.15]	1.63[1.02-2.61]	1.66[1.01-2.71]	1.57[0.98-2.52]	1.43[0.93-2.20]	1.23[0.89-1.72]	1	0	0
871-Plumbers and pipe fitters	29/22	1.94[1.03-3.65]	1.62[0.95-2.77]	1.90[1.06-3.43]	1.87[1.05-3.32]	1.82[1.05-3.16]	1.76[1.04-2.98]	2	0	0
891-Glass formers, cutters, grinders and finishers	15/8	1.95[0.78-4.85]	1.46[0.75-2.87]	1.53[0.72-3.24]	1.40[0.71-2.78]	1.24[0.70-2.19]	1.06[0.71-1.57]	0	1	0
751-Fibre preparers	5/5	1.97[0.48-8.03]	1.29[0.58-2.89]	1.31[0.47-3.64]	1.18[0.48-2.90]	1.05[0.52-2.12]	0.94[0.61-1.46]	0	0	0
943-Non-metallic mineral product makers	7/4	2.00[0.56-7.13]	1.34[0.61-2.92]	1.51[0.64-3.56]	1.42[0.68-2.98]	1.31[0.73-2.35]	1.21[0.82-1.78]	0	1	1
723-Metal melters and reheaters	12/6	2.06[0.75-5.72]	1.45[0.71-2.96]	1.84[0.87-3.90]	1.79[0.91-3.50]	1.73[0.99-3.03]	1.70[1.11-2.61]	1	2	0
791-Tailors and dressmakers	14/11	2.08[0.87-5.00]	1.54[0.79-2.97]	1.64[0.77-3.50]	1.49[0.74-3.01]	1.28[0.71-2.32]	1.05[0.70-1.57]	0	0	0
893-Glass and ceramics kilnmen	11/6	2.14[0.73-6.25]	1.45[0.70-3.01]	1.61[0.76-3.42]	1.53[0.79-2.95]	1.44[0.85-2.45]	1.38[0.95-2.00]	1	1	1
796-Upholsterers and related workers	19/11	2.27[0.99-5.21]	1.65[0.87-3.13]	1.79[0.87-3.69]	1.62[0.82-3.18]	1.37[0.77-2.45]	1.08[0.72-1.62]	0	0	0
728-Metal platers and coaters	13/7	3.26[1.17-9.07]	1.81[0.88-3.72]	2.30[1.05-5.05]	2.08[1.02-4.23]	1.82[1.00-3.29]	1.57[1.01-2.46]	0	2	0

ISCO CODE - OCCUPATION	CASES / CONTROLS	ML	SB	HR				CARCINOGENIC EXPOSURE		
			7-fold range	**τ = 0.76** *(20-fold range)*	**τ = 0.59** *(10-fold range)*	**τ = 0.41** *(5-fold range)*	**τ = 0.23** *(2.5-fold range)*			
		OR[95%CI]	**OR[95%CI]**	**OR[95%CI]**	**OR[95%CI]**	**OR[95%CI]**	**OR[95%CI]**	**ASB[a]**	**CR[a]**	**SI[a]**
612-Specialised farmers	9/3	3.44[0.90-13.17]	1.59[0.71-3.55]	1.81[0.67-4.93]	1.53[0.63-3.68]	1.23[0.61-2.47]	1.00[0.65-1.55]	0	0	0
982-Ships' engine-room ratings	8/2	5.88[0.94-36.71]	1.54[0.64-3.73]	2.43[0.79-7.46]	2.16[0.82-5.73]	1.93[0.87-4.29]	1.78[0.96-3.32]	2	0	0

[a] ASB=Asbestos (0=no exposure, 1=low exposure, 2=high exposure), CR=Chromium (0=no exposure, 1=low exposure, 2=high exposure), SI=Silica (0=no exposure, 1=low exposure, 2=high exposure)

Discussion

In our analyses, HR provided estimates which are likely to be more reliable and have narrower confidence intervals than are obtained with conventional ML analysis. Many of the more extreme estimates obtained through ML analysis are based on small numbers and have large confidence intervals. HR, by including prior information on exposure to three lung carcinogens in a second-stage model, pulls these estimates towards their respective prior means and thereby reduces their estimated standard errors and confidence intervals. The strength and direction of the shrinkage for the more extreme estimates depend on the prior estimated exposures of the corresponding occupations to the three carcinogens. For example, ‘‘specialised farmers’’ are not exposed to any of the considered carcinogens and HR therefore pulls the corresponding OR strongly towards the null value whereas the OR remains elevated for ‘‘metal melters and reheaters’’ who are exposed to both asbestos and chromium. In a situation of multiple comparisons, HR is thus a useful tool for data analysis which takes into account the multiple comparisons involved and the commonalities of exposures across different occupations.

In our analyses, HR and SB shrinkage had similar effects on the ML estimates. However, since HR uses more detailed prior information than SB, the shrinkage performed by the former method is likely to be more appropriate and specific than the latter (provided of course that this prior information is reasonably valid). Our findings show that all the estimates were shrunk towards the null value through SB whereas some of them were pulled in the opposite direction by HR, because of the use of additional prior information. Thus, both approaches aim at decreasing false-positive findings but HR also mitigates the inherent effect of the shrinkage of increasing false-

negatives. On the other hand, SB is easier to compute and does not need the manipulation of a second-stage matrix. The choice between the two methods therefore essentially depends on the availability and the reliability of the information included in the second-stage model.

The HR shrinkage as proposed in this paper could have two relevant implications when conducting exploratory analyses on risks associated with occupations: i) it decreases the possibility that an occupation entailing exposure to important known occupational carcinogens is dismissed by the study, ii) it helps to pick up, among occupations not entailing exposure to known occupational carcinogens, those which should be further investigated and are more likely to provide information on the role of new or suspected occupational carcinogens. Our findings on construction painters, which were associated with an OR of 1.85 (95% CI: [1.0-3.15]) in the standard ML approach, are an example of the latter implication. According to the DOM-JEM construction painters are not exposed to chromium or silica and have a low exposure to asbestos. However, the OR remains elevated after HR even when using a τ of 0.23 (OR = 1.23, 95% CI: [0.8-1.72]), suggesting that any increased risk is due to other exposures. Thus it is worth conducting further studies on painters. Indeed a recent meta-analysis on 47 independent estimates of the association between employment as a painter and risk of lung cancer estimated an overall relative risk of 1.35 (95% CI: [1.2-1.41]), which is closer to our HR than our ML estimate[187]. If HR weighs information from the DOM-JEM too heavily, we might incur in the problem that high risks for occupations classified as unexposed to the 3 considered carcinogens (but likely to be exposed to other carcinogens) are always knocked down. Among the 20 occupations with the highest ML ORs, 6 were unexposed to asbestos, chromium or silica. HR shrinkage was strong for risks based on a small

number of subjects, but did not nullify those based on larger numbers, such as upholsterers (ML OR: 2.27, HR OR [τ = 0.59]: 1.62) and tailors/dressmakers (ML OR: 2.08, HR OR [τ = 0.59]: 1.49).

The inclusion of many covariates in the second-stage model can lead to collinearity problems and difficulties in estimating second-stage coefficients. For this reason, our analyses were restricted to three well known lung carcinogens from the DOM-JEM[185]. The JEM used here classifies the exposure to the carcinogens in three levels, and these were used to specify the second-stage model. Before fitting the model, we verified that a sufficient number of subjects were exposed to each level of the selected carcinogens to ensure model convergence. If this condition did not hold, a simpler version of the matrix with dichotomous exposure to the carcinogens could have been used. An interesting future development of this method could be the use of continuous exposure variables in the second-stage model.

In our analyses, we have assessed the impact of four different values of τ. The choice of τ depends on how many second-stage covariates are included in the model, how strong and reliable their associations with both the outcome and the exposures of interest are, and how well the first-stage model was specified (i.e. if it can be assumed that all the relevant confounders have been included). In our analyses, we chose to include three well known strong occupational lung carcinogens, and our first-stage model was adjusted for smoking. It was therefore reasonable to assume that 95% of the estimates would lie within a maximum 10-fold-range of each other (e.g. between 0.5 and 5.0) after accounting for the second-stage covariates, and a τ of 0.59 would then be appropriate. For each occupation, τ was inversely weighted by the amount of exposure

to carcinogens as specified in the JEM. In this respect, HR is superior to SB since it modulates the weights given to the residual variation of each occupation and hence the amount of shrinkage towards prior information.

HR has already been shown to be a valid approach to adjust for multiple comparisons in studies involving the analysis of multiple occupational exposures and outcomes[183] and in occupational studies where the first-stage exposures (chemical and physical agents) were regressed on physicochemical properties in a second-stage model[51, 58]. In our analyses, we focused on the risks associated with the occupations and included carcinogens in a second-stage model. We found that HR could also be a valuable tool in occupational studies in which the risk of disease is estimated for a large amount of occupations when we have information available on the key carcinogenic exposures involved in each occupation. With the constant progress in exposure assessment methods in occupational settings and the construction and refinement of Job Exposure Matrices, it should become easier to have access to this information and carry out this type of analysis in the future.

Acknowledgments

We are grateful to D. Mirabelli for his helpful advice.

SECTION 3

Systematic error

CHAPTER VI

Adjustment for exposure misclassification – Application of several methods in a case-control study of lung cancer where the smoking status has been misclassified

Information bias in epidemiologic studies arises from measurement error[188]. For discrete variables, measurement error is more usually termed ‘classification error’ or ‘misclassification’. Such errors can occur either at the stage of the study design (e.g. invalid or inadequate questions or proposed methods for biomarker analyses or environmental exposure measurements), or at the stage of data collection (e.g. information not accurately reported or recorded, errors due to imperfect measurement techniques and instruments), or subsequently (e.g. mistakes in data entry)[20].

Misclassification can occur for exposure, disease status of the study subject and/or confounders. Also, misclassification can be non-differential (i.e. when misclassification of exposure does not depend on the status of the subject with respect to other variables in the analysis, including disease[188]) or differential (i.e. when misclassification of exposure depends on disease status or on other variables in the analysis). In the current

chapter we focus on the example of non-differential misclassification but the methods applied here can also be employed when misclassification is differential.

As discussed in Chapter II, measurement error can be considered as a missing-data problem[62] in that information has been recorded on a variable which is an imperfect surrogate measurement for the 'true' variable of interest. When internal validation data is available, both the misclassified and true values are known for a subsample of the dataset and the true values for the misclassified variable are only partially missing. When internal validation or replication data are not available, the true values for the mismeasured variables are completely missing.

In this chapter we consider the very common situation where exposure has been misclassified and potential confounders (or matching factors) need to be adjusted for while estimating the exposure-disease association. We explore four different approaches: multiple imputation for measurement error (MIME), imputation based on specifying the sensitivity and specificity (SS), Direct Imputation (DI) of the 'true' exposure using a regression model for the predictive values, and imputation based on a fully Bayesian analysis to correct for exposure misclassification[6]. These four methods are applied to estimate the association between smoking status and lung cancer in simulated data obtained by misclassifying the smoking status data from the study analysed in Chapter III[3]. We only illustrate here the case of non-differential misclassification i.e. we assume that misclassification is independent of the other variables (disease status, gender). However, these methods can also be applied when misclassification is differential.

We first consider the case where validation data is available for a subsample of the study participants. In this first situation, Multiple Imputation for Measurement Error[74] (MIME) can be used to impute the missing data for the true exposure, given the misclassified exposure, confounders, outcome and other covariates. We have applied this approach, considering three different sizes of validation subsamples.

We considered the association between smoking status (ever/never) and lung cancer[6]. The odds ratio (OR) of lung cancer for ever-smokers vs. never-smokers was estimated using unconditional logistic regression, adjusting for gender. The SAS Logistic procedure (SAS V9.3) was used to estimate ORs and corresponding 95% confidence intervals (95%CI).

We assumed that our original dataset was correctly specified, i.e. that the "true" smoking status T was known for all subjects, and we then misclassified T with a sensitivity of 0.8 and a specificity of 0.9. Let *Y*, *C*, and *X* denote the indicators for case-control status, gender (1=Man, 0=Woman) and classified smoking status, and let n_{tycx} denote the number of subjects with $T=t$, $Y=y$, $C=c$, and $X=x$. To create the misclassified smoking status X, we computed the frequencies n_{tyc+} in each of the 8 combinations of the categories of *T*, *Y* and C, where a subscript "+" indicates summation over a subscript. We then calculated the frequencies of classified ever/never smokers n_{tycx} for each of these combinations as follows:

$$n_{1yc1} = n_{1yc+} \times 0.8$$
$$n_{1yc0} = n_{1yc+} \times 0.2$$
$$n_{0yc1} = n_{0yc+} \times 0.1$$
$$n_{0yc0} = n_{0yc+} \times 0.9.$$

In the original dataset, three validation subsamples in which both X and T were assumed to be known for a percentage of subjects - approximately 15%, 25% and 50% of the study population - were randomly extracted; the 50% subsample included the 25% subsample, which included the 15% subsample. The OR of lung cancer for ever-smokers vs. never-smokers was then 'corrected' using the two validation subsamples using Multiple Imputation for Measurement Error (MIME)[69, 74].

The association between the true and misclassified smoking status was estimated through the following unconditional logistic regression model in the validation subsample:

$$\text{logit}\left[P(T=1|X=x,Y=y,C=c)\right]=\beta_0+\beta_X x+\beta_Y y+\beta_C c+\beta_{YC} yc$$

where the *xc* and *xy* product terms are omitted as the nondifferentiality assumption implies that the association between *T* and *X* does not depend on *Y* and *C* and therefore $\beta_{XY}=\beta_{XC}=0$.

The MIME algorithm involved the following steps[74]:

i. For m iterations (Here we chose m=40 in a way that kept the simulation error small[189])
    a. Imputation of T for the subjects not included in the validation subsample, based on the associations estimated in the validation subsample
    b. Estimation of the log odds ratio of lung cancer for ever-smoking using the corrected smoking status: $\log OR_{TY}$
ii. Combination of the 40 $\log OR_{TY}$ estimated using Rubin's rules[71]

The “true” odds ratio of lung cancer for ever-smokers vs. never-smokers adjusted for sex in the original dataset was OR=8.18 95%CI [5.86-11.43] (log odds ratio (logOR)=2.10 95%CI [1.77-2.44])[6]. After misclassification it decreased to OR=3.08 95%CI [2.40-3.96] (logOR=1.13 95%CI [0.87-1.38]).

Table VI.1 shows the odds ratios (ORs) and 95% confidence intervals (CIs) of lung cancer after the application of MIME. After applying MIME correction with our three randomly selected validation subsamples, the OR of lung cancer for ever-smokers vs. never-smokers was strongly pulled towards the true OR. The confidence interval, which had become much narrower after misclassification, became larger than the original confidence interval after MIME adjustment, thus reflecting the uncertainty due to both random sampling and misclassification. The bigger the validation subsample, the closer the adjusted OR was to the true OR. The true logOR was always included in the 95% confidence interval, even with a validation subsample size of only 15%.

Table VI.1. Odds ratios of lung cancer and respective 95% CIs after the application of MIME

Subsample size	OR	LCL	UCL
15%	6.18	3.08	12.40
25%	6.91	3.83	12.45
50%	7.60	5.02	11.51

The following paper now considers the case where validation data is not available.

A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable

Marine Corbin, Stephen Haslett, Neil Pearce, Milena Maule and Sander Greenland

Background: Measurement error is an important source of bias in epidemiological studies, which can be considered as a missing-data problem.

Methods: We consider three approaches to sensitivity analysis: Imputation of the 'true' exposure based on specifying the sensitivity and specificity of the measured exposure (SS), Direct Imputation (DI) using a regression model for the predictive values, and imputation based on a fully Bayesian analysis. We apply these three approaches in a case-control study of lung cancer where the main exposure, the smoking status, has been deliberately misclassified after data collection. The first two approaches (but not the third as it is always probability-based) are implemented using fixed-parameter (FBA) and probabilistic (PBA) bias analyses.

Results: The "true" smoking-lung cancer odds ratio (OR) adjusted for sex in the original dataset was OR=8.18 (95% confidence limits (CL): 5.86-11.43); after misclassification it decreased to OR=3.08 (95%CL: 2.40-3.96). The ORs estimated from all three approaches were always closer to the true OR than the OR estimated with the misclassified smoking status. The intervals obtained from the three approaches were wider than the confidence interval obtained with the misclassified smoking status. When the misclassification parameters were misspecified, the true OR was often omitted in the FBA intervals whereas it was always included in the PBA and Bayesian intervals.

Conclusion: These results support recommendations for PBA and Bayesian analyses when a risk assessment must be made that accounts for all sources of uncertainty.

- This manuscript was submitted for publication. -

Introduction

A major source of bias and uncertainty in epidemiologic analysis is measurement error, usually termed “misclassification” when referring to discrete variables[20, 188]. As with all kinds of error, measurement error can be considered as a missing-data problem[62] in that information has been recorded on a variable which is an imperfect surrogate for the ‘true’ variable of interest (for which the information is missing). When internal validation data are available, both the misclassified and true values are known for a subsample of the dataset; thus true values for the misclassified variable are only partially missing, and standard methods for missing data, including multiple imputation, data augmentation, the “Expectation-Maximisation” (EM) algorithm and inverse probability weighting, can be applied to correct for misclassification[63-65, 67-70, 74, 190]. Similar methods can be used when replication data are available[190].

When internal validation or replication data are not available, however, the true values for the mismeasured variables are completely missing and the relations among them are not statistically identified, in that no consistent point estimate can be constructed from the data without adding further, potentially arbitrary assumptions. Thus, standard missing-data software cannot be used without the addition of pseudo-validation data representing external (prior) information relating true values to observed variables[62]. Ideally, this analysis should be repeated using different plausible priors. Simpler sensitivity analysis formulae can be used to adjust for misclassification assuming fixed misclassification rates which can be based on the literature or on external validation data[20, 191-194].

In this paper we focus on the situation where exposure has been misclassified, no validation data are available, and adjustment for potential confounders or matching factors is needed. We present applications of different methods to adjust for the misclassification of smoking status in a case-control study of smoking and lung cancer, while also adjusting for gender. Each method can be carried out with commercial software. Our aim is to illustrate each method by applying them to real data, with known added misclassification structure and thus compare their strengths and limitations.

Material and methods

We consider three different approaches to imputation:

1. Imputation of the 'true' exposure based on specifying the sensitivity and specificity (SS) of the measured exposure with respect to the 'true' exposure, where sensitivity is the probability a truly exposed person is classified as exposed, and specificity is the probability a truly unexposed person is classified as unexposed.
2. Direct imputation (DI) of the 'true' exposure using a regression model for the predictive values, where the positive predictive value is the probability a person classified as exposed is truly exposed and the negative predictive value is the probability a person classified as unexposed is truly unexposed. These predictive values can then be used to impute the true exposure values.
3. Imputation of the 'true' exposure based on a fully Bayesian analysis. This analysis is intended to:
    - Establish the links to the two previous approaches

- Assess when the two previous approaches provide a useful approximation to a fully Bayesian analysis

For SS and DI, we start with the simplest case where the relevant parameters (sensitivity and specificity for SS and regression model coefficients for DI) are fixed, i.e. fixed-parameter bias-sensitivity analysis (FBA), before conducting a probabilistic bias-sensitivity analysis (PBA). In FBA, an *a priori* set of several combinations of values for the parameters is specified and an analysis is then conducted for each combination to see how sensitive the misclassification adjustments are to the values chosen. In PBA, one specifies prior probability distributions for the parameters and incorporates those into the analysis process.

The classical way to do PBA is via Bayesian techniques[75], but a simple approximation is afforded by Monte Carlo sensitivity analysis (MCSA) in which combinations of parameters are sampled from the prior distributions, and then an analysis is done for each sampled combination. Thus, MCSA is sensitivity analysis using a random sample of values for adjustments, instead of fixed values. On the other hand, a fully Bayesian analysis updates the prior distributions based on the data to yield posterior distributions for the parameters[20, 191, 195, 196]. Procedures for MCSA have been implemented in Excel and SAS[20, 195]. These programs do not take covariates into account in the imputation process, although it is possible to do so[196].

Here, we use updated versions of a SAS macro implementing MCSA, which allow covariates in the imputation model[195]. We then compare the results obtained with MCSA with those obtained with fully Bayesian analyses implemented with the free software WinBUGS. We have chosen the parameterisations and parameter values

(either as prior distributions or as fixed values depending on the method) with the intention of enabling direct comparisons between the three methods.

We assume that misclassification is nondifferential, i.e. independent of the other variables (disease status, gender). The methods applied here can however be employed when misclassification is differential by expanding the models[20, 62, 195, 196].

Description of the data

The data pertain to a population-based lung cancer case-control study conducted in New Zealand[3]. Briefly, cases were all subjects diagnosed with incident lung cancer notified to the New Zealand Cancer Registry during 2007 and 2008 and aged 20-75 years. Controls were recruited from the New Zealand Electoral Rolls of 2003 and 2008 and were frequency matched with the cases for age and gender. The study was performed according to the Declaration of Helsinki and was approved by the Multiregion Ethics Committee (AKL/99/172). Informed consents were obtained from all study subjects.

We considered the association between smoking status (ever/never) and lung cancer. The odds ratio (OR) of lung cancer for being ever-smoker vs. never-smoker was estimated using unconditional logistic regression, adjusting for gender. The SAS Logistic procedure (SAS V9.3) was used to estimate ORs and corresponding 95% confidence intervals (95%CI).

We assumed that our original dataset was correctly specified, i.e. that the "true" smoking status indicator T was known for all subjects. Our focus was on comparing the three methods, and we therefore chose misclassification values which enabled these

comparisons, but we also attempted to use realistic misclassification rates which had been observed in previous studies. In 9 studies using the cotinine validation method reported by a meta-analysis[197], the lowest sensitivity of the self-reported smoking status was 0.82 and the lowest specificity was 0.91. We therefore misclassified T with a sensitivity of 0.8 and a specificity of 0.9. Let *Y*, *C*, and *X* denote the indicators for case-control status, gender (1=Man, 0=Woman) and misclassified smoking status, and let n_{tycx} denote the number of subjects with *T=t*, *Y=y*, *C=c*, and *X=x*. To create the misclassified smoking status X, we computed the frequencies n_{tyc+} in each of the 8 combinations of the categories of *T*, *Y* and C, where a subscript "+" indicates summation over a subscript. We then calculated the frequencies of classified ever/never smokers n_{tycx} for each of these combinations as follows:

$$n_{1yc1} = n_{1yc+} \times 0.8$$
$$n_{1yc0} = n_{1yc+} \times 0.2$$
$$n_{0yc1} = n_{0yc+} \times 0.1$$
$$n_{0yc0} = n_{0yc+} \times 0.9.$$

1. Sensitivity/Specificity Imputation Analysis (SS)

A. Fixed-parameter bias-sensitivity analysis (SS FBA)

Initial sensitivity and specificity values se^0 and sp^0 were specified and the associations between the misclassified smoking status and the other variables were estimated by fitting the following unconditional logistic regression model to the data:

$$\text{logit}\left[P\left(X=1 \middle| Y=y, C=c\right)\right] = \gamma_0 + \gamma_Y y + \gamma_C c + \gamma_{YC} yc \qquad (1)$$

FBA was then applied according to the following steps:

i. Estimation of $\pi^* = P(X=1)$ for each individual using the estimates of model (1) coefficients

ii. Restriction of se^0 and sp^0 according to the following equations in order to confine $\pi = P(T=1)$ to values between 0 and 1

$se = \max(se^0, \hat{\pi}^*)$; $sp = \max(sp^0, 1-\hat{\pi}^*)$

where $\hat{\pi}^*$ are the estimates of π^* for each individual obtained in i.

iii. Calculation of the positive predictive values (PPV) and the negative predictive values (NPV) according to the following equations

$$PPV = \frac{se \times (\hat{\pi}^* + sp - 1)}{\hat{\pi}^* \times (se + sp - 1)}; \; NPV = \frac{sp \times (se - \hat{\pi}^*)}{(1-\hat{\pi}^*)(se + sp - 1)}$$

iv. Calculation of the frequencies of subjects in each strata $T \times Y \times C$ from PPV and NPV

v. Estimation of the C-adjusted ln OR_{TY}

vi. Use of the jackknife procedure to account for the uncertainty in the estimation of $\hat{\pi}^*$ (reiteration of steps i. to v. for each 'leave one out' sample from the original data) and calculation of the 95% confidence interval (CI) for ln OR_{TY} from the jackknife standard error (SE).

Initial sensitivity and specificity values se^0 and sp^0 were set to combinations of the following values: se^0 (0.7, 0.8, 0.9) and sp^0 (0.8, 0.9, 1).

B. Probabilistic bias-sensitivity analysis (SS PBA)

We assumed that self-reported smoking status was better than chance in terms of prediction of the true smoking status[76] (i.e. sensitivity and specificity were both greater than 0.5). Logit-transformed scaled normal prior distributions were therefore

specified for sensitivity and specificity so that both parameters would fall in the interval [0.5,1].

We defined

$$Se^0 = 0.5 + 0.5\text{expit}(\lambda)$$
$$Sp^0 = 0.5 + 0.5\text{expit}(\varepsilon)$$

with normal prior distributions on λ and ε, as specified in Table VI.2. In order to determine the parameters for these prior distributions, we first chose 95% limits for sensitivity and specificity, converted these limits into limits for λ and ε by solving the above equations, and calculated prior means and prior standard deviations for λ and ε from these limits.

The association between misclassified smoking status and the other variables was then estimated by fitting again model (1) to the data.

PBA was implemented via the following MCSA algorithm:

i. For 10,000 iterations

   a. Estimation of $\pi^* = P(X=1)$ for each individual using the estimates of model (1) coefficients

   b. Random draw of se^0 and sp^0 from their prior distributions

   c. Restriction of se^0 and sp^0: $se = \max(se^0, \hat{\pi}^*)$; $sp = \max(sp^0, 1-\hat{\pi}^*)$ where $\hat{\pi}^*$ are the estimates of π^* for each individual obtained in a.

   d. Calculation of PPV and NPV for each individual

   e. Calculation of the frequencies of subjects in each strata $T \times Y \times C$ from PPV and NPV

   f. Estimation of the C-adjusted ln OR_{TY}

g. Use of the jackknife procedure to account for the uncertainty in the estimation of π^* (reiteration of steps a. to f. for each 'leave one out' sample from the original data) and calculation of the jackknife standard error of ln OR_{TY}.

h. Perturbation of ln OR_{TY} with its jackknife standard error: $\ln\widetilde{OR}_{TY}$ = random draw from a normal distribution with mean= ln OR_{TY} and standard deviation = jackknife SE(ln OR_{TY})

ii. Computation of the mean, median, and 2.5th and 97.5th percentiles from the distribution of the 10,000 ln OR_{TY} and $\ln\widetilde{OR}_{TY}$ estimates, and their antilogs. We refer to the resulting 2.5th and 97.5th simulation percentiles for OR_{TY} and $\widetilde{OR}_{TY}$ as 95% simulation limits (SL) for OR_{TY}, under the given (possibly truncated) priors.

Table VI.2. Prior distributions on sensitivity and specificity for SS PBA

	Prior values mean (standard deviation)		Means [95% limits] for sensitivity and specificity	
Set of priors	λ	ε	Se^0	Sp^0
1	-0.41(0.5)	0.41(0.5)	0.7[0.60,0.82]	0.8[0.68,0.90]
2	0.41(0.5)	1.39(0.5)	0.8[0.68,0.90]	0.9[0.80,0.96]
3	0.41(1.5)	1.39(1.5)	0.8[0.54,0.98]	0.9[0.59,0.99]
4	1.39(0.5)	3.89(0.5)	0.9[0.80,0.96]	0.99[0.97,1.00[

2. Direct Imputation Analysis (DI)

A. Fixed-parameter bias-sensitivity analysis (DI FBA)

The probability of being a true ever-smoker $P(T = 1)$ was estimated from a logistic regression (model (2)).

$$\text{logit}\left[P(T=1|X=x,Y=y,C=c)\right] = \beta_0 + \beta_X x + \beta_Y y + \beta_C c + \beta_{YC} yc \qquad (2)$$

where the *xc* and *xy* product terms are omitted because they are zero under nondifferential misclassification.

We gave fixed values to all model (2) coefficients as shown in Table VI.3. The values were chosen on the basis of published data and surveys and our prior assumptions about sensitivity, specificity and ln OR_{TY}. Values for sensitivity, specificity, ln OR_{TY} and prevalences of true smokers in strata of *Y* and *C* were translated into values for model (2) coefficients using the equations in Table VI.4. Details of the calculations are available in Appendix AII.2. Unlike sensitivity and specificity, model (2) coefficients can be any real number.

Table VI.3. Fixed values for model (2) coefficients in DI FBA

	Fixed values for model (2) coefficients					Values for sensitivity, specificity, OR_{TY}				Values for prevalence of T=1	
Set of values	β_0	β_X [a]	β_Y [a]	β_C [a]	β_{YC} [a]	Sensitivity	specificity	$OR_{TY}(C=0)$	$OR_{TY}(C=1)/ OR_{TY}(C=0)$	P(T=1|Y=0,C=0)	P(T=1|Y=0,C=1)
1	-1.37	2.23	1.94	0.10	0.46	0.7	0.8	6.93	1.59	0.403	0.428
2	-1.90	3.58	1.94	0.10	0.46	0.8	0.9	6.93	1.59	0.403	0.428
3	-2.69	6.79	1.94	0.10	0.46	0.9	0.99	6.93	1.59	0.403	0.428
4	-1.37	2.23	1.25	0.10	0.46	0.7	0.8	3.5	1.59	0.403	0.428
5	-1.90	3.58	1.25	0.10	0.46	0.8	0.9	3.5	1.59	0.403	0.428
6	-2.69	6.79	1.25	0.10	0.46	0.9	0.99	3.5	1.59	0.403	0.428
7	-1.37	2.23	2.64	0.10	0.46	0.7	0.8	14	1.59	0.403	0.428
8	-1.90	3.58	2.64	0.10	0.46	0.8	0.9	14	1.59	0.403	0.428
9	-2.69	6.79	2.64	0.10	0.46	0.9	0.99	14	1.59	0.403	0.428

[a] *X*=misclassified smoking status; *Y*=case/control status; *C*=gender

Table VI.4. Definition of model (2) coefficients for DI FBA

Coefficient	Definition
β_0	$\text{expit}(\beta_0) = \dfrac{\exp(\beta_0)}{1+\exp(\beta_0)}$ $= P(T=1 \mid X=0, Y=0, C=0)$ $= \dfrac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0) + P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}$ $= \dfrac{\left(\dfrac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0)}\right)}{\left(1+\dfrac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0)}\right)}$ Then, $\exp(\beta_0) = \dfrac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0)}$ $= \dfrac{(1-se_{00})P(T=1 \mid Y=0, C=0)}{sp_{00}(1-P(T=1 \mid Y=0, C=0))}$ $= \dfrac{(1-se)P(T=1 \mid Y=0, C=0)}{sp(1-P(T=1 \mid Y=0, C=0))}$ as we are assuming nondifferentiality

Cont.

Coefficient	Definition
β_X	$\exp(\beta_X) = \dfrac{\left[\dfrac{P(T=1\mid X=1,Y=y,C=c)}{P(T=0\mid X=1,Y=y,C=c)}\right]}{\left[\dfrac{P(T=1\mid X=0,Y=y,C=c)}{P(T=0\mid X=0,Y=y,C=c)}\right]} = \dfrac{\left[\dfrac{P(X=1\mid T=1,Y=y,C=c)}{P(X=0\mid T=1,Y=y,C=c)}\right]}{\left[\dfrac{P(X=1\mid T=0,Y=y,C=c)}{P(X=0\mid T=0,Y=y,C=c)}\right]} = \dfrac{se_{yc}\times sp_{yc}}{(1-se_{yc})\times(1-sp_{yc})} = \dfrac{se\times sp}{(1-se)\times(1-sp)}$, as we are assuming nondifferentiality
β_Y	$\exp(\beta_Y) = \mathrm{OR}_{TY}(X=x,C=0) = \mathrm{OR}_{TY}(C=0)$, as we are assuming nondifferentiality
β_C	$\exp(\beta_C) = \dfrac{\dfrac{P(T=1\mid C=1,X=x,Y=0)}{P(T=0\mid C=1,X=x,Y=0)}}{\dfrac{P(T=1\mid C=0,X=x,Y=0)}{P(T=0\mid C=0,X=x,Y=0)}} = OR_{TC}(X=x,Y=0) = OR_{TC}(Y=0)$, as we are assuming nondifferentiality
β_{YC}	$\exp(\beta_{YC}) = \dfrac{\left\{\left[\dfrac{P(T=1\mid X=x,Y=1,C=1)}{P(T=0\mid X=x,Y=1,C=1)}\right]\right\}}{\left\{\left[\dfrac{P(T=1\mid X=x,Y=0,C=1)}{P(T=0\mid X=x,Y=0,C=1)}\right]\right\}} = \dfrac{\left\{\left[\dfrac{P(Y=1\mid T=1,X=x,C=1)}{P(Y=0\mid T=1,X=x,C=1)}\right]\right\}}{\left\{\left[\dfrac{P(Y=1\mid T=0,X=x,C=1)}{P(Y=0\mid T=0,X=x,C=1)}\right]\right\}} = \dfrac{OR_{TY}(X=x,C=1)}{OR_{TY}(X=x,C=0)}$, $= \dfrac{OR_{TY}(C=1)}{OR_{TY}(C=0)}$, as we are assuming nondifferentiality

The following algorithm was then applied:

i. Computation of $\hat{\pi}$, the estimate of $P(T=1)$ from model (2) and calculation of PPV and NPV for each individual

ii. Calculation of the frequencies of subjects in each strata $T \times Y \times C$ from PPV and NPV

iii. Estimation of the C-adjusted ln OR_{TY} and 95% CI.

B. Probabilistic bias-sensitivity analysis (DI PBA)

The probability of being a true ever-smoker $P(T = 1)$ was represented by model (2).

We then placed normal prior distributions on all model (2) coefficients as shown in Table VI.5. Means and 95% limits for sensitivity, specificity, ln OR_{TY} and prevalences of true smokers in strata of Y and C were translated into prior means and standard deviations for model (2) coefficients using the equations in Table VI.4. In order to allow the comparison between DI PBA, SS PBA and the fully Bayesian analysis described next, prior means, standard deviations and correlation for coefficients β_0 and β_X were estimated by simulation (see Appendix AII.2 for details). If DI PBA were used alone with no intent to compare it with other methods, parameters for the prior distributions of β_0 and β_X could be specified directly based on background information, without simulation. Appendix AII.2 provides an approximate estimate of the correlation ρ_{0X} between β_0 and β_X.

Table VI.5. Prior distributions on model (2) coefficients for DI PBA

Set of priors	Prior values						Means [95% limits] for sensitivity, specificity, OR_{TY}, OR_{TC}, and prevalence of T=1					
	mean (standard deviation)					Correlation (β_0 , β_X)						
	β_0	β_{X} [(a)]	β_{Y} [(a)]	β_{C} [(a)]	β_{YC} [(a)]	ρ_{0X}	sensitivity	specificity	$OR_{TY}(C=0)$	$OR_{TC}(Y=0)$	$OR_{TY}(C=1)/OR_{TY}(C=0)$	P(T=1|Y=0,C=0)
1	-1.39 (0.23)	2.28 (0.47)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.80	0.7[0.60,0.82]	0.8[0.68,0.90]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
2	-1.92 (0.31)	3.64 (0.57)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.72	0.8[0.68,0.90]	0.9[0.80,0.96]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
3	-2.08 (0.93)	3.93 (1.66)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.74	0.8[0.54,0.98]	0.9[0.59,0.99]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
4	-2.71 (0.40)	6.81 (0.66)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.65	0.9[0.80,0.96]	0.99[0.97,1.00[	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
5	-1.92 (0.31)	3.64 (0.57)	1.25 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.72	0.8[0.68,0.90]	0.9[0.80,0.96]	3.5 [0.89,13.76]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
6	-1.92 (0.31)	3.64 (0.57)	2.64 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.72	0.8[0.68,0.90]	0.9[0.80,0.96]	14[3.55,55.26]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]

[(a)] *X*=misclassified smoking status; *Y*=case/control status; *C*=gender

The following MCSA algorithm was then applied:

i. For 100,000 iterations

   a. Random draw of model (2) coefficients from their respective prior distributions

   b. Computation of $\hat{\pi}$, the estimate of $P(T=1)$ from model (2) for each individual

   c. Imputation of *T* from a Bernoulli distribution with probability of success $\hat{\pi}$

   d. Computation of a C-adjusted ln OR_{TY} from the imputed TYC data

   e. Perturbation of ln OR_{TY} with the original XY random error: $\ln \tilde{O}R_{TY}$ = random draw from a normal distribution with mean= ln OR_{TY} and standard deviation = SE(ln OR_{XY})

ii. Computation of the mean, median, and 2.5th and 97.5th percentiles from the distribution of the 100,000 ln OR_{TY} and $\ln \tilde{O}R_{TY}$ estimates, and their antilogs.

3. Fully Bayesian analysis

A fully Bayesian analysis was then applied, in which prior distributions were chosen to allow direct comparison between the three methods. The model was modified from Chu et al.'s[16] to include the gender *C*.

We specified prior distributions for two groups of parameters:

a) The sensitivity and specificity, defining the association between the true smoking status T and the misclassified smoking status *X*

As in SS PBA, we defined:

$$Se^0 = 0.5 + 0.5\text{expit}(\lambda)$$
$$Sp^0 = 0.5 + 0.5\text{expit}(\varepsilon)$$

and we placed normal prior distributions on λ and ε.

b) The association of T with case/control status Y and gender C

The prevalence of true smokers in the population was defined as a function of *Y* and *C* (model (3)).

$$\text{logit}\left[P\left(T=1\middle|Y=y,C=c\right)\right]=\alpha_0+\alpha_Y y+\alpha_C c+\alpha_{YC} yc \qquad (3)$$

We placed normal prior distributions on all model (3) coefficients. Values for the parameters of these prior distributions were obtained by giving means and 95% limits to ln OR_{TY} and prevalences of true smokers in strata of *Y* and *C*, and by converting those into prior means and standard deviations for α_0, α_Y, α_C, α_{YC} using the equations in Table VI.6. Unlike model (2) in DI, model (3) did not include the misclassified smoking status *X* as prior distributions were already specified for the association between *T* and *X* in a). Therefore, while DI model (2) coefficients β_0, β_C, β_Y and β_{YC} are functions of *X*, *Y* and *C*, model (3) coefficients α_0, α_Y, α_C, α_{YC} only depend on *Y* and *C*. However, we saw in Table VI.4 that since we are considering nondifferential misclassification, the associations between *T* and *Y* and between *T* and *C* do not depend on *X* and hence

$$\beta_Y=\log\left(OR_{TY}\left(X=x,C=0\right)\right)=\log\left(OR_{TY}\left(C=0\right)\right)=\alpha_Y$$

$$\beta_C=\log\left(OR_{TC}\left(X=x,Y=0\right)\right)=\log\left(OR_{TC}\left(Y=0\right)\right)=\alpha_C$$

$$\begin{aligned}\beta_{YC}&=\log\left(OR_{TY}\left(X=x,C=1\right)\right)-\log\left(OR_{TY}\left(X=x,C=0\right)\right)\\&=\log\left(OR_{TY}\left(C=1\right)\right)-\log\left(OR_{TY}\left(C=0\right)\right)\\&=\alpha_{YC}\end{aligned}$$

Markov Chain Monte Carlo (MCMC) in WinBUGS was used to sample from the posterior distribution. Two Markov chains were run using the block Gibbs sampler with 800,000 iterations following 10,000 discarded for burn-in.

In the first MCMC analysis, in order to allow direct comparison with SS PBA, we placed the same informative prior distributions on λ and ε as in SS PBA (see Table VI.2) while we placed vague prior distributions on α_0, α_Y, α_C and α_{YC}, as specified in Table VI.7a.

In the second MCMC analysis, in order to allow direct comparison with DI PBA (see Table VI.5), we placed informative distributions on all parameters as specified in Table VI.7b. The simulation linking the prior distributions for coefficients λ, ε and α_0 to the prior distributions for DI PBA coefficients β_0 and β_X is described in Appendix AII.2. We then placed the same prior distributions on α_Y, α_C, α_{YC} as on β_Y, β_C, β_{YC}, respectively (Table VI.7b).

SAS and WinBUGS codes are available in Appendix III.

Table VI.6. Definition of model (3) coefficients for the fully Bayesian analysis

Coefficient	Definition
α_0	$\exp it(\alpha_0) = \dfrac{\exp(\alpha_0)}{1+\exp(\alpha_0)} = P(T=1 \mid Y=0, C=0)$
α_Y	$\exp(\alpha_Y) = \mathrm{OR}_{TY}(C=0)$
α_C	$\exp(\alpha_C) = OR_{TC}(Y=0)$
α_{YC}	$\exp(\alpha_{YC}) = \dfrac{\left\{\left[\dfrac{P(T=1\mid Y=1,C=1)}{P(T=0\mid Y=1,C=1)}\right]\right\}}{\left\{\left[\dfrac{P(T=1\mid Y=0,C=1)}{P(T=0\mid Y=0,C=1)}\right]\right\}} = \dfrac{\left\{\left[\dfrac{P(Y=1\mid T=1,C=1)}{P(Y=0\mid T=1,C=1)}\right]\right\}}{\left\{\left[\dfrac{P(Y=1\mid T=0,C=1)}{P(Y=0\mid T=0,C=1)}\right]\right\}} = \dfrac{OR_{TY}(C=1)}{OR_{TY}(C=0)},$

Table VI.7a. Prior distributions for the fully Bayesian analysis corresponding to the SS PBA analysis (Table VI.2)

	Prior values mean (standard deviation)						Means [95% limits] for sensitivity, specificity, OR_{TY}, OR_{TC}, and prevalence of T=1					
Set of priors	λ	ε	α_0	α_Y	α_C	α_{YC}	sensitivity	specificity	$OR_{TY}(C=0)$	$OR_{TC}(Y=0)$	$OR_{TY}(C=1)/ OR_{TY}(C=0)$	P(T=1|Y=0,C=0)
1	-0.41(0.5)	0.41(0.5)	0(2)	0(1.5)	0(4)	0(3)	0.7[0.60,0.82]	0.8[0.68,0.90]	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
2	0.41(0.5)	1.39(0.5)	0(2)	0(1.5)	0(4)	0(3)	0.8[0.68,0.90]	0.9[0.80,0.96]	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
3	0.41(1.5)	1.39(1.5)	0(2)	0(1.5)	0(4)	0(3)	0.8[0.54,0.98]	0.9[0.59,0.99]	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
4	1.39(0.5)	3.89(0.5)	0(2)	0(1.5)	0(4)	0(3)	0.9[0.80,0.96]	0.99[0.97,1.00[	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]

Table VI.7b. Prior distributions for the fully Bayesian analysis corresponding to the DI PBA analysis (Table VI.5)

	Prior values mean (standard deviation)						Means [95% limits] for sensitivity, specificity, OR_{TY}, OR_{TC}, and prevalence of T=1					
Set of priors	λ	ε	α_0	α_Y	α_C	α_{YC}	sensitivity	specificity	$OR_{TY}(C=0)$	$OR_{TC}(Y=0)$	$OR_{TY}(C=1)/ OR_{TY}(C=0)$	P(T=1|Y=0,C=0)
1	-0.41(0.5)	0.41(0.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.7[0.60,0.82]	0.8[0.68,0.90]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
2	0.41(0.5)	1.39(0.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.68,0.90]	0.9[0.80,0.96]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
3	0.41(1.5)	1.39(1.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.54,0.98]	0.9[0.59,0.99]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
4	1.39(0.5)	3.89(0.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.9[0.80,0.96]	0.99[0.97,1.00[	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
5	0.41(0.5)	1.39(0.5)	-0.39(0.07)	1.25(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.68,0.90]	0.9[0.80,0.96]	3.50[0.89,13.76]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
6	0.41(0.5)	1.39(0.5)	-0.39(0.07)	2.64(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.68,0.90]	0.9[0.80,0.96]	14.00[3.55,55.26]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]

Results

The "true" odds ratio of lung cancer for ever-smokers vs. never-smokers adjusted for sex in the original dataset was OR=8.18 (95% CL 5.86,11.43) (log odds ratio (ln OR)=2.10, 95% CL 1.77, 2.44). After misclassifying the smoking status with a sensitivity of 0.8 and a specificity of 0.9, it decreased to OR=3.08, 95% CL 2.40, 3.96 (ln OR=1.13, 95%CL 0.87, 1.38).

A preliminary analysis was conducted in order to check what possible values of sensitivity and specificity could have led to the misclassified odds ratio[198].

Let π_{YC} be the prevalence of subjects truly ever-smokers and π_{YC}^{*} the prevalence of subjects classified as ever-smokers in the different strata of *Y* and *C*. Then

$$\pi_{YC} = \frac{\pi_{YC}^{*} + Sp - 1}{Se + Sp - 1}$$

The prevalences π_{YC} are restricted to values between 0 and 1, which implies the following restrictions:

If $Se + Sp > 1$

$Se > \max_{YC}\left(\pi_{YC}^{*}\right)$ and $Sp > \max_{YC}\left(1 - \pi_{YC}^{*}\right)$

If $Se + Sp < 1$

$Se < \min_{YC}\left(\pi_{YC}^{*}\right)$ and $Sp < \min_{YC}\left(1 - \pi_{YC}^{*}\right)$

Table VI.8 shows the prevalences of subjects classified as ever-smokers π_{YC}^{*} and never-smokers $1 - \pi_{YC}^{*}$ in strata of *Y* and *C*. Therefore, the restrictions on Se and Sp become:

If $Se + Sp > 1$

$Se > \pi_{11}^{*}$ (i.e. $Se > 0.76$) and $Sp > 1 - \pi_{00}^{*}$ (i.e. $Sp > 0.59$)

If $Se + Sp < 1$

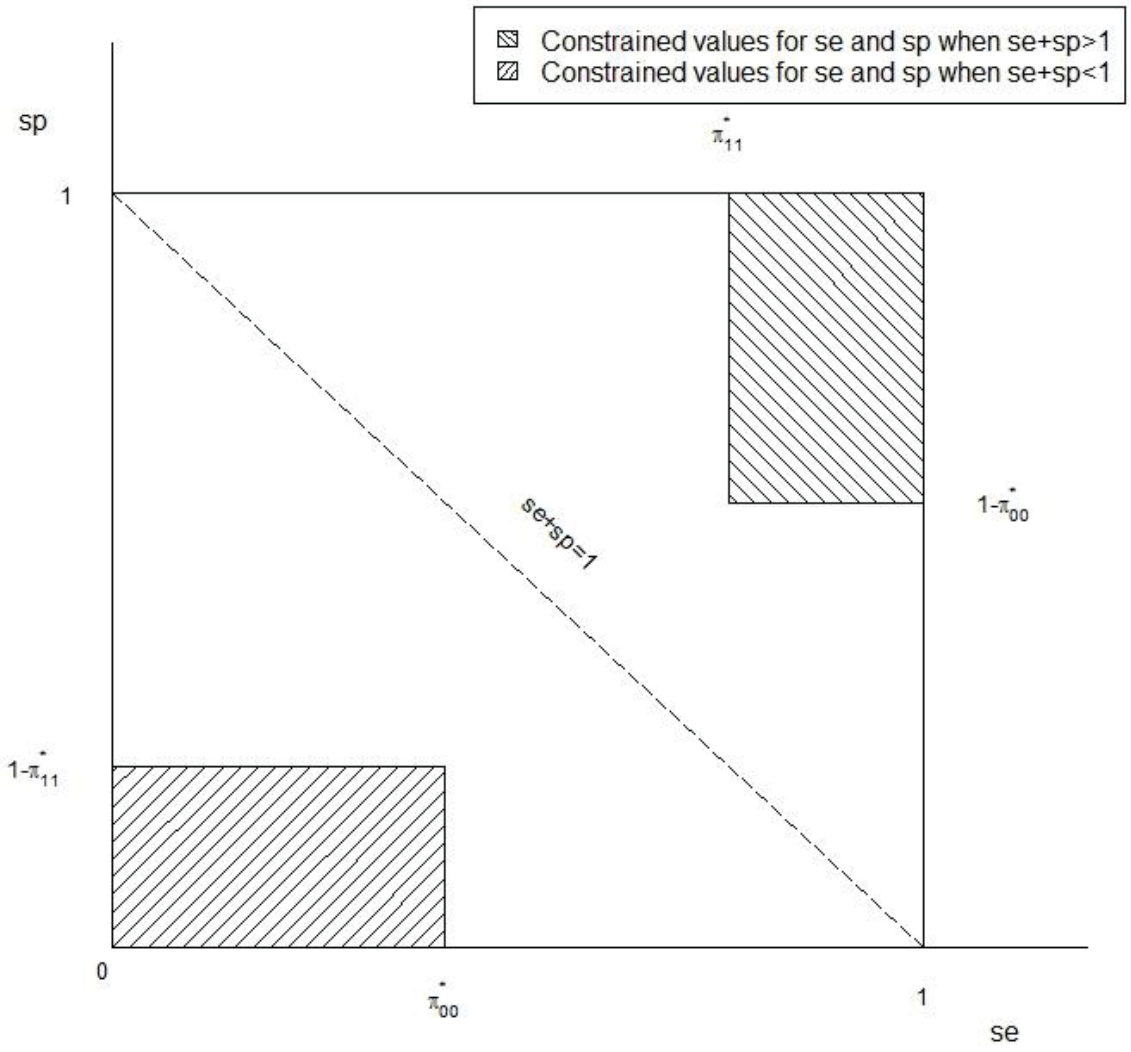
$Se < \pi_{00}^{*}$ (i.e. $Se < 0.41$) and $Sp < 1 - \pi_{11}^{*}$ (i.e. $Sp < 0.24$)

The possible values for the sensitivity and specificity which respect these restrictions in our data are represented in Figure VI.1. As we assumed that self-reported smoking status was better than chance, we only considered the case where $Se + Sp > 1$.

Table VI.8. Prevalences of subjects classified as exposed and non-exposed in strata of *Y* and *C*

Y	*C*	π_{YC}^{*}	$1 - \pi_{YC}^{*}$
0	0	0.41	0.59
0	1	0.52	0.48
1	0	0.69	0.31
1	1	0.76	0.24

Figure VI.1. Description of possible ranges for misclassification parameters



Tables VI.9 to VI.14 below contain results for each of the methods separately.

Tables VI.9 and VI.10 show the smoking-lung cancer odds ratios from SS FBA and DI FBA, respectively. As expected, for larger values of se^0 and sp^0, the OR obtained with SS FBA became closer to the OR obtained with the misclassified smoking status. The OR estimate appeared more sensitive to changes in the sensitivity than in the specificity of the measured exposure. When se^0 was 0.7, the sensitivity was replaced by 0.77 in step ii. of the algorithm.

Similarly, DI FBA produced adjusted ORs closer to the OR obtained with the misclassified smoking status when sensitivity and specificity were assumed to be higher (i.e. when β_X was given a higher value). However, the adjusted OR was more sensitive to the prior mean given to β_Y than to the prior means given to β_0 and β_X. When $\exp(\beta_0) = 0.15$ and $\exp(\beta_X) = 36$, i.e. when the sensitivity and the specificity were assumed to equal the actual sensitivity and specificity of the misclassified exposure, the ln OR obtained with DI FBA was very close to the value given to β_Y.

For both SS FBA and DI FBA, the confidence intervals obtained after adjustment were wider than the intervals obtained with the 'standard' analysis using misclassified smoking status. The intervals became narrower on the logarithmic scale when increasing the sensitivity and specificity and when decreasing β_Y for DI FBA. The intervals were wider when using SS FBA than when using DI FBA, as SS FBA also attempted to account for the uncertainty in estimating π^*.

Table VI.9. Smoking-lung cancer odds ratios from SS FBA

$se^0 \setminus sp^0$	**0.8**	**0.9**	**1**
0.7	15.67 [5.51,44.60]	13.00 [4.77,35.40]	11.59 [4.35,30.89]
0.8	9.54 [4.75,19.16]	8.00 [4.15,15.43]	7.18 [3.80,13.54]
0.9	5.15 [3.47,7.63]	4.38 [3.07,6.26]	3.96 [2.83,5.54]

Table VI.10. Smoking-lung cancer odds ratios from DI FBA

$exp(\beta_0,\beta_X)$ \ $exp(\beta_Y)$	3.5	6.93	14
(0.25,9.33)	5.40 [4.10,7.11]	9.56 [6.94,13.17]	17.72 [11.94,26.30]
(0.15,36)	4.86 [3.69,6.39]	7.60 [5.59,10.33]	12.89 [8.96,18.54]
(0.07,891)	3.88 [2.97,5.07]	5.06 [3.81,6.72]	7.40 [5.41,10.13]

Tables VI.11 and Table VI.12 show the smoking-lung cancer odds ratios from SS PBA and DI PBA, respectively. In SS PBA, out of 10,000 draws of initial sensitivity se^0, 8,799 (88%), 3,128 (31%), 4,283 (43%) and 76 (0.76%) values for prior distributions 1, 2, 3 and 4, respectively, were lower or equal to 0.76 and were adjusted to 0.77. In draws of initial specificity sp^0 from prior distribution 3, 415 (4%) were lower or equal to 0.59 and were adjusted to 0.60. An increase of the prior means for the parameters λ and ε (i.e. an increase of the means for the sensitivity and specificity) resulted in a decrease of the geometric mean and median ORs. As expected, prior distributions 2 and 3, assigning prior means for the sensitivity and the specificity equal to the actual misclassification sensitivity and specificity gave the closest geometric mean and median ORs to the true OR. When increasing the 95% limits for the sensitivity and specificity, the geometric mean and median ORs were slightly increased, stepping away from the true OR and the 95% simulation intervals (95% SI) were much wider. In DI PBA, assigning prior means for the sensitivity and the specificity equal to the actual misclassification sensitivity and specificity, and a prior mean for β_Y equal to the true ln OR of lung cancer in women (sets of priors 2 and 3) gave the closest geometric mean and median ORs to the true OR. As for DI FBA, an increase in absolute values of the prior means for β_0 and β_X (set of priors 4) still resulted in a decrease of the geometric mean and median ORs, while increasing the prior mean for β_Y (set of priors 6) considerably increased the mean and median ORs. Increasing the prior standard error of

β_X (i.e. increasing the 95% limits for sensitivity and specificity, set of priors 3), slightly increased the geometric mean and median ORs and the 95% SI. For both SS and DI, the PBA approach yielded simulation intervals which were much wider than the confidence intervals obtained with the original and the misclassified smoking status. Adjustment for random error modified more the findings in SS PBA than in DI PBA. This is because adjustment for random error also accounted for the uncertainty in the estimation of π^* in SS PBA.

Table VI.11. Smoking-lung cancer odds ratios from SS PBA

Set of priors [a]	Incorporation of estimated random error[b]	GeoMean OR	Median OR	LSL	USL
1	No	15.40	15.33	8.05	25.53
	Yes	15.36	15.31	4.78	48.20
2	No	8.24	7.95	4.38	14.54
	Yes	8.24	8.18	2.99	23.17
3	No	8.97	10.86	3.37	49.46
	Yes	8.93	9.05	2.15	63.86
4	No	4.21	4.00	3.39	7.23
	Yes	4.22	4.12	2.64	8.09

[a] described in Table VI.2
[b] step i.h. of SS PBA algorithm

Table VI.12. Smoking-lung cancer odds ratios from DI PBA

Set of priors [a]	Incorporation of estimated random error[b]	GeoMean OR	Median OR	LSL	USL
1	No	9.58	9.32	2.97	35.92
	Yes	9.57	9.33	2.89	36.90
2	No	7.83	7.44	3.17	25.57
	Yes	7.82	7.45	3.06	26.07
3	No	7.38	6.78	3.10	27.72
	Yes	7.37	6.81	2.93	28.26
4	No	5.38	5.02	3.15	13.52
	Yes	5.38	5.07	2.92	13.94
5	No	4.98	4.81	2.20	13.71
	Yes	4.98	4.83	2.11	14.09
6	No	13.22	12.59	4.59	50.74
	Yes	13.22	12.59	4.45	52.00

[a] described in Table VI.5
[b] step i.e. of DI PBA algorithm

Tables VI.13 and VI.14 show smoking-lung cancer odds ratios from MCMC analyses 1 and 2, respectively.

As with SS PBA, median ORs obtained from MCMC analysis 1 decreased when increasing the parameters λ and ε. However, median ORs obtained from MCMC analysis 1 were higher than median ORs obtained with SS PBA. Ninety-five percent credibility intervals (95%CI) obtained from MCMC analysis 1 were also wider than the 95% SI obtained with SS PBA, suggesting that SS PBA underestimates the uncertainty in the prevalence of true smokers in strata of *T* and *Y*.

In comparison with median ORs obtained from DI PBA, median ORs obtained from MCMC analysis 2 were more sensitive to the prior means assigned to the sensitivity and specificity and less sensitive to the prior mean assigned to $OR_{TY}(C=0)$ (i.e. $\exp(\beta_Y)$ in DI PBA and $\exp(\alpha_Y)$ in MCMC analysis 2). Ninety-five percent credibility intervals (95%CI) obtained from MCMC analysis 2 were slightly narrower than DI PBA 95% SI.

When the means assigned to sensitivity and specificity equalled the actual misclassification sensitivity and specificity (sets of priors 2 and 3), the informative prior distributions placed on Model 3 coefficients in MCMC analysis 2 yielded median ORs closer to the true OR than in MCMC analysis 1. Ninety-five percent credibility intervals (95%CI) were narrower after MCMC analysis 2 than after MCMC analysis 1.

Table VI.13. Smoking-lung cancer odds ratios from MCMC analysis 1

Set of priors	Median OR	LCL	UCL
1	18.5	6.72	86.24
2	11.63	4.44	48.51
3	12.08	3.60	64.51
4	4.75	2.93	18.48

Table VI.14. Smoking-lung cancer odds ratios from MCMC analysis 2

Set of priors	Median OR	LCL	UCL
1	12.10	5.67	35.38
2	8.03	4.23	21.78
3	7.09	3.28	23.50
4	4.47	3.14	8.12
5	6.70	3.85	15.40
6	10.63	4.83	37.09

Discussion

In this paper we have illustrated the use of several currently available methods for bias analysis, which can be implemented using standard statistical software. Although these methods have been applied previously[196, 199], to our knowledge they have not yet been compared and contrasted in parallel analyses as we have done here.

Sensitivity/specificity (SS) imputation analysis has the advantage of requiring only the specification of *a priori* values for sensitivities and specificities. When one wishes to

incorporate uncertainty about these values, one can specify prior distributions and then sample values from those. Prior distributions for sensitivity and specificity seem intuitive but require restriction to the range of values compatible with the data (some values may be impossible given the observed data)[75]. Furthermore, the apparent simplicity of the SS approach has its own difficulties, since seemingly intuitive guesses for sensitivity and specificity may turn out to be highly implausible when compared to what one might deduce by considering the actual classification mechanism and background literature, particularly when covariates are also taken into account.

In Direction Imputation analysis (DI) , we directly modelled predictive values, an approach which eliminates concerns about constraints on sensitivity and specificity[62]. The main limitations of this approach, however, are that the user needs to specify values or prior distributions for coefficients about which there may be poor prior information, and that the resulting adjusted estimate can also be quite sensitive to the prior mean specified for the association of interest (here, the odds ratio of lung cancer for being ever smoker).

Both SS and DI methods have been applied using fixed-parameter bias-sensitivity analysis (FBA) and probabilistic bias-sensitivity analysis (PBA). FBA is simpler and faster to run, since one only needs to specify fixed values. It is also very useful to check which values are compatible with the data in the SS method. Nonetheless, it does not account for uncertainty in the specification of the bias parameters. PBA takes this uncertainty into account and produces estimates with large simulation intervals, which tend to be less sensitive to misspecification of the bias parameters than the estimates obtained after FBA.

Adjustment for uncertainty due to random error is also allowed by PBA. Adjustment for random error in the XY association requires only the addition of a random number to estimates during simulation. This shortcut thus leads to fast run times, but should be used with caution as it may seriously underestimate the actual contribution of random error to uncertainty about TY; this underestimation will be a problem if uncertainty due to random error is not minor compared to uncertainty about the classification parameters. Bootstrap or jacknife methods for adding random error are preferable, but can lead to long run times; bootstrapping in particular can encounter technical problems in small samples[200].

The choice between SS and DI depends on what information is available. In particular, one needs to evaluate the amount and the quality of prior information to decide between whether to set priors on sensitivity and specificity or on regression coefficients for predictive values. When both validation data and prior information are available, all the information can be combined using data augmentation[2, 62, 186, 201], in which prior distributions are translated into new data records and added to the validation data. Such an approach enables analysis with standard methods for missing data.

Bayesian procedures may be preferable to PBA, especially when one feels comfortable assigning priors to parameters beyond the classification model[75]. Our MCMC analyses indicated that the uncertainty in the prevalence of exposure might be underestimated when using SS PBA. In addition, unlike SS PBA, MCMC analyses do not require truncation of the prior distributions when the sensitivity or specificity prior extends below the range compatible with the data as the prior distributions are updated

based on the data to yield posterior distributions. For further analysis and contrast of SS PBA and Bayesian analyses, see Maclehose and Gustafson[75].

It has been remarked that most epidemiologists write their methods and results sections as frequentists and their introduction and discussion sections as Bayesians[1, 62]. In their methods and results sections, they analyse their data as if those are the only data that exist, and as if there is no bias left uncontrolled by covariate adjustment (i.e. they implicitly use dogmatic point-null priors on hidden bias parameters[62]). In the discussion, they then assess their results relative to background information, examining consistency with previous studies, biological plausibility, and the possibility of various biases. It has been lamented however that the latter discussions severely overweight their own results, and tend to understate biases in these results, displaying especially poor intuitions about potential misclassification effects[20, 191, 196].

These problems can be ameliorated by including bias analyses[20, 62, 191, 202]. We have reviewed and illustrated several methods feasible using standard statistical software. Hopefully, sensitivity and bias analyses will become options in standard statistical packages to supplement existing methods, facilitating conduct and presentation of bias analysis before inferences are offered. This will enable readers to better quantitatively assess problems of bias in drawing conclusions from the report. FBA is particularly simple and may be useful for initial bias analyses, but we recommend PBA or Bayesian analyses when doing a risk assessment that must account for all sources of uncertainty.

Acknowledgements

This project was funded by the Health Research Council of New Zealand, by the New Zealand Department of Labour, by Lottery Health Research, by the Cancer Society of New Zealand, and by the Accident Compensation Corporation (ACC). The Centre for Public Health Research is supported by a Programme Grant from the Health Research Council. We wish to thank Jonathan Bartlett for his comments on earlier analyses which led to the production of this paper.

Conflict of interest

The authors declare no conflict of interest.

SECTION 4

Discussion and conclusions

CHAPTER VII

General discussion

This thesis has explored the application of several Bayesian approaches, implemented with standard statistical software, in environmental and occupational epidemiology. These methods have been applied to several different studies of occupational risks for lung cancer and upper aerodigestive tract cancer, and the findings are therefore of interest in themselves. However, the focus of the thesis has on the application of Bayesian methods to produce these findings. It has not been intended to represent a comprehensive overview of all possible Bayesian methods, but rather as an exploration of the Bayesian methods which are most relevant and appropriate for the occupational cancer case-control studies which I have analysed.

The main findings are summarised below, followed by a discussion of methodological issues, limitations, and recommendations for further research

A. Key findings in occupational epidemiology of lung cancer and upper aerodigestive tract cancer

Chapters III, IV and V identified several professional categories associated with increased risks of lung and upper aerodigestive tract (UADT) cancer.

a) Metal workers

In both occupational studies of lung cancer in New Zealand and Italy (respectively Chapters III and V), several occupations involving work with metal were found to be associated with an increased risk of lung cancer. These included different categories of metal processing such as metal drawers and extruders, metal melters and reheaters, metal platers and coaters; they also included other job categories such as welders and flame cutters, and plumbers and pipe fitters. According to the Job Exposure Matrix (DOM-JEM) used in Chapter V, most of these occupations entailed exposure to chromium and a few of them entailed exposure to asbestos or silica.

b) Machinery operators

An increased risk of lung cancer was observed for plant and machine operators and assemblers, and for the sub-categories of electric and electronic equipment assemblers and machine tool operators in the New Zealand study. In the Italian study, the categories of farm machinery operators and ships' engine-room ratings, highly exposed to silica and asbestos respectively, had an increased risk of lung cancer.

c) Other manufacturing workers

Workers involved in the manufacturing of non-metallic minerals had increased risks of lung cancer in both studies. The DOM-JEM identified a low exposure to chromium and silica for this job category. In the New Zealand study, an increased risk was also found

for other rubber and plastics products machine operators and for subjects employed in the petroleum, coal, chemical and associated product manufacturing industry. In the Italian study, glass and ceramic kiln workers - who were exposed to low levels of asbestos, chromium and silica - had an increased risk of lung cancer.

d) Construction workers

Builder's labourers had an increased risk of lung cancer in the New Zealand study. In the Italian study, some categories of construction labourers (construction painters and reinforced-concreters, cement finishers and terrazzo workers) involving low exposure to asbestos, chromium or silica had an increased risk of lung cancer. Several categories of construction workers including concreters, bricklayers, painters and workers in the construction of roads or the erection of roofs were also found to have an increased risk of UADT cancer in the European case-control study in Chapter IV.

e) Textile workers

An increased risk of lung cancer was found for textile products machine operators in New Zealand with a duration-response association. An elevated risk was also observed for tailors and dressmakers in both the New Zealand and Italian studies, and for upholsterers and related workers in the Italian study only. These two professions apparently did not involve any exposure to asbestos, chromium or silica.

f) Drivers

Drivers, in particular heavy truck drivers, had an increased risk of lung cancer in the New Zealand study, with a positive duration-response association. In the Italian study,

hierarchical regression indicated an elevated risk for railway engine drivers and firemen, who were highly exposed to asbestos.

g) Wood workers

Wood processing and papermaking plant operators had an increased risk of lung cancer in the New Zealand study, with a particularly strong association for timber processing machine operators that increased with the duration of employment. An elevated risk of lung cancer was estimated for loggers in the same study. Loggers were also found to be have an increased risk of UADT cancer, in particular when the employment lasted at least ten years.

h) Other occupations

The New Zealand study also found increased risks of lung cancer for meat workers and health care professionals. However, these occupations had no increased risk in the Italian study. An increased risk of UADT cancer was observed for several categories of dairy workers. This association has not been reported previously and requires further investigation.

Semi-Bayes adjustments towards the global mean and hierarchical regression helped identifying, among all occupations and industries showing an elevated risk of cancer, which ones would be worth investigating further. This selection was done by considering the variation in both point estimates and 95% confidence intervals after SB adjustment or HR. In Chapter III, several *a priori* high risk occupations (loggers, pressers, electric and electronic equipment assemblers) which were not associated with a statistically significant increased risk had an OR estimate greater than 2 after SB

adjustment. These occupations were highlighted by SB adjustment while a conventional analysis would not have considered them for further investigation. In addition, SB adjustment performed a selection among all occupations associated with a statistically significant increased risk in the conventional analysis and drew a particular attention to some occupations and industries such as nursing associate professionals, enrolled nurses, construction labourers, car retailing, road freight transport and psychiatric hospitals. In Chapter IV, for the occupations/industries associated with a statistically significant increased risk of upper aerodigestive tract cancer, the OR estimates were all between 1.31 and 1.90 after SB adjustment. For some of these occupations/industries associated with large frequencies (bricklayers, lorry and van drivers, labourers, general construction of buildings and civil engineering works) , the OR estimates and 95%CI remained almost unchanged by SB adjustment, indicating that the estimated increased risks were unlikely to be due to chance and worth investigating further. On the other hand, some industries associated with a statistically significant increased risk of UADT cancer (other retail sale in non-specialised stores, mining of uranium and thorium ores) saw their OR estimate and their 95%CI pulled down by SB adjustment, suggesting that these findings should be interpreted with precaution. In Chapter V, most of the occupations associated with the twenty highest OR estimates would have been neglected in the conventional analysis as the estimates were not statistically significant. HR highlighted the occupations which were worth investigating further because of their exposure to carcinogens (e.g. metal drawers and extruders, metal melters and reheaters). In addition, HR also identified occupations which involved no or little exposure to known carcinogens but for which the OR estimate and 95%CI did not vary much after SB adjustment (e.g. painters) and which might then involve exposure to other carcinogens.

B. Bayesian methods to account for random error

1. Summary of the approach

In the New Zealand and Italian case-control studies of lung cancer, and in similar studies elsewhere, typically we may have a small group of occupations and exposures which are of a priori interest (e.g. asbestos, silica, etc), but we may also collect information on literally hundreds of other occupations and exposures for which there is little or no prior evidence and we are just 'having a look' to see what we find. If we test hundreds of occupations, and none of them are really causes of lung cancer, then we will get a false positive finding (i.e. a significantly increased or decreased risk) about 5% of the time, e.g. in a study of 500 occupations we might expect to have 25 false positive findings by chance. The classic solution to this problem is the Bonferroni correction[18], which adopts more stringent criteria for deciding which p-values will be considered 'significant' based on the number of tests that we have done (and/or adjusts the confidence intervals for each effect estimate to reduce the probability that any individual effect estimate will not include the null value).

The problem with this approach is that in most real life situations it does not work well. Suppose we do a case-control study of asbestos and lung cancer, and along the way we take a complete occupational history, and therefore we are able to 'have a look' at the associations of 500 other occupations with lung cancer risk. Should we 'adjust' our asbestos findings because we have also looked at 500 other occupations? Surely, asbestos was the main point of doing the study, and there is strong prior evidence that asbestos causes lung cancer. Thus, the finding of an association of asbestos exposure with lung cancer is in a completely different league than, for example, if we also tested 500 other occupations and found that teachers had a higher risk of lung cancer. For each

occupation/exposure, the key issue is not how many other associations we tested, but what the prior evidence was before we did the study (very strong in the case of asbestos, non-existing in the case of teachers), and what further evidence the new study adds.

Thus, if we repeated the study in a new population, we would expect the asbestos effect to be replicated, whereas chance associations (e.g. teachers) would be likely to exhibit 'regression to the mean' and would usually not show such strongly increased risks again. Bayesian methods can be used to take this likely regression to the mean into account, and to 'shrink' our effect estimates based on how strong the regression to the mean is likely to be. We can do this shrinkage for different groups of exposures. For example, in a case-control study of asbestos and lung cancer, the occupational exposures might fall into three groups: (i) asbestos, the main exposure under study; (ii) other exposures/occupations for which there was a priori evidence that they were risk factors for lung cancer; (iii) all other occupations/exposures. For group I (asbestos), there is no need to shrink the estimated odds ratio, since asbestos essentially forms its own group. For group III (other occupations/exposures), the estimated odds ratio for each occupation is shrunk towards the overall mean for these occupations (which will usually be close to 1.0). For group II, if we consider it reasonable to assume that these 'a priori high risk occupations' have similar odds ratios (e.g. in previous studies they have generally been found to all have odds ratios of about 2.0), then it is reasonable to consider these occupations as a group. The estimated odds ratio for each occupation is shrunk towards the overall mean for these high risk occupations.

In this thesis, I have illustrated the use of shrinkage methods with case-control studies from New Zealand and Italy.

2. Key findings

a) Either semi-Bayes (SB) adjustments or hierarchical regression are needed to account for random error in epidemiological studies involving the estimation of multiple associations.

In Chapters III, IV and V, we reported that a number of occupations/industries were associated with extreme odds ratios (OR) estimates. Those estimates sometimes truly represented the presence of a strong association between the occupation/industry with lung or UADT cancer. However, in other cases, extreme estimates were only the results of random error (i.e. false positive findings). Adjustment for multiple comparisons was then necessary to identify which occupations/industries were actually worth investigating further. SB adjustment towards the global mean and hierarchical regression were appropriate in this situation because they allowed to take *a priori* belief into account and shrunk the point estimates towards their prior means so that the shrinkage was stronger for less reliable estimates based on small frequencies. For example, in Chapter III, enrolled nurses and draughting technicians had both very high OR estimates before SB adjustment. However, while enrolled nurses were represented by 25 cases and 7 controls, draughting technicians were represented only by 4 cases and 3 controls and the shrinkage of the estimate by SB adjustment resulted to be much stronger for the second occupation than for the first one. Similarly, in Chapter V, the shrinkage operated by SB adjustment towards the global mean and hierarchical regression was much stronger for glass formers, cutters, grinders and finishers than for plumbers and pipe fitters, represented by higher numbers.

b) Hierarchical regression may be superior to SB adjustment when reliable second-stage information is available.

In Chapter V, we observed that SB adjustment towards the global mean systematically shrunk the OR estimates towards 1 and that this shrinkage was always inversely related to the number of subjects involved in the considered occupation. Therefore, occupations represented by few subjects in our data always had their OR estimate strongly pulled towards 1 by this approach although some of these occupations actually entailed exposure to lung carcinogens. Hence, by decreasing the number of false positive findings, SB adjustment towards the global mean revealed some false negative findings. In Chapter III, SB adjustment within the finer groups of *a priori* high risk occupations and industries shrunk some of the OR estimates upwards away from 1. Hierarchical regression uses a second-stage model which allows us to specify a different prior mean for each estimate. The shrinkage is still (roughly) inversely related to the number of subjects in a particular category, but depending on the value of the prior mean, it can be directed either towards or away from 1. For example, in Chapter V, metal annealers, temperers and case-hardeners were highly exposed to chromium, a well-known lung carcinogen. Their OR estimate before any adjustment for multiple comparisons was 1.14. SB adjustment towards the global mean shrunk this estimate downwards to 1.08 as only 4 cases and 7 controls were employed in this occupation. On the other hand, hierarchical regression pulled this estimate upwards as the exposure to a lung carcinogen yielded a high prior mean.

3. Limitations

a) Grouping the estimates

A major assumption of Semi-Bayes (SB) and Empirical Bayes (EB) adjustments is that the estimates being shrunk towards a same value have to be "exchangeable". This means that they must be considered as arising from a single distribution around the same mean. It is difficult to group the estimates according to this criterion when carrying out an exploratory analysis if very little information is known about which exposures are more likely to be at risk than others. However, if the 'exposures' included are only occupations and industries, it is unlikely that some of them would be associated with extremely high or low risks compared with the others.

Moreover, when applying SB adjustment, we created a priori homogeneous ensembles to shrink towards a group mean. In Chapters III, IV and V, we grouped the estimates according to the degree of specification of the occupation/industry using the number of digits of the codes provided by the jobs/industries classifications. In addition, in Chapter III, we included the occupations/industries which were assumed to be *a priori* at risk according to previous studies in a separate group. In Chapter IV and Chapter V, occupations were divided into two main subsets: blue-collar worker and white-collar worker occupations.

Hierarchical regression groups the estimates using a second-stage model. This allows a more specific division and a more appropriate shrinkage provided the second-stage covariates are chosen appropriately and well-informed. In Chapter V, the difficulty stemmed from the choice of a Job Exposure Matrix (JEM) which would suit our data. This means that the selected carcinogens had to be associated with an increased risk of

lung cancer and that their levels of exposure had to be known for each occupation in our data. Another subtle issue to address while applying hierarchical regression was the collinearity between the second-stage covariates. In Chapter V, we restricted the number of lung carcinogens to three in the second-stage model in order to overcome this problem.

b) Determination of the prior variance

A second difficulty lies in determining the weights to be given to prior information. In EB adjustments, these weights are computed from the data. However, we noted in Chapter II that, although this method has the advantage of not requiring the specification of any prior value, the calculation of EB estimates often yields negative variances and becomes then impossible.

SB adjustments and hierarchical regression require the specification of prior values. Choosing prior variances is a delicate issue: variances which are too large would correspond to uninformative priors and yield insufficient shrinkage; on the other hand, variances which are too small may over-shrink estimates and neglect the information contained in the observed data. In SB adjustment, the weight given to prior information is computed from the variance of the true parameters around their group mean. The prior variance is better chosen according to findings of previous studies on a similar subject. In chapters III, IV and V, we set the prior variance at 0.25, assuming, as suggested previously[48], that 95% of the relative risks should lie within a 7-fold range of each other in an occupational context.

In hierarchical regression, the weights given to prior information vary with the estimates and the specification of prior values becomes then more complex. The weights are obtained from the second-stage variance-covariance matrix, which states the variation of the estimates around their respective prior means. In Chapter V, the variation of each estimate corresponded to the variance of the residual effect of the occupation after taking into account the effects of the three lung carcinogens. For the analyses, we assumed that each variance was a function of two components. The first component was common for all variances and it was used to control the global strength of the shrinkage. In order to check the sensitivity of the results to this first component, we set it to four different values. The second component varied across the estimates and the main difficulty was to select the factors it had to vary with, so that valid information would be added while keeping the calculations simple enough. We assumed that the more second-stage covariates we had, the less likely it would be to have a large residual association[53]. We therefore set this component so that the more carcinogenic exposures were involved in an occupation and the stronger these exposures were, the smaller the residual variation and the bigger the weight given to prior information would be.

C. Bayesian methods to adjust for systematic error

1. Summary of the approach

Systematic errors include selection bias, information bias and (residual) confounding. In this thesis I have focussed on applying methods for the assessment and the correction of information bias, but similar methods can be applied for selection bias and confounding[62].

Information bias in epidemiologic studies arises from measurement error, which can occur at several stages (study design, data collection, data entry)[20]. Here I have presented a case-control study of lung cancer, where the exposure of interest, the smoking status, has been misclassified. I have considered two situations:

1) Validation data is available: The true values for the smoking status are only partially missing. In this situation I have applied Multiple Imputation for Measurement Error (MIME) to adjust for misclassification assuming three different sizes of validation substudy.
2) Validation data is not available: The true values for the smoking status are totally missing. In this situation I have explored three different methods to conduct sensitivity analysis: Imputation based on specifying the sensitivity and specificity (SS) of the measured smoking status with respect to the 'true' unknown smoking status, Direct imputation (DI) of the 'true' smoking status using a regression model for the predictive values and imputation based on a fully Bayesian analysis.

2. Key findings

a) When one has access to a minimum of valid information (either internal or external to the study) about the association between the mismeasured/misclassified variable and the true variable, it is recommended to adjust the estimate for measurement error in the statistical analyses.

In Chapter VI, we saw that, in both situations 1) and 2), the odds ratio of lung cancer for ever-smoker vs. never smokers was closer to the 'true' odds ratio after adjustment than when using the misclassified smoking status. Table VII.1 shows the bias obtained with the misclassified smoking status and after all adjustment methods where the bias was

computed as the estimated log odds ratio minus the 'true' log odds ratio. Even in situation 2), when no validation data was available and wrong prior information was used, the resulting bias was smaller in absolute value than the bias obtained with the misclassified smoking status.

Table VII.1. Bias in log odds ratio estimated in Chapter VI with the misclassified smoking status (naïve) and after adjustment using MIME, SS Fixed-parameter Bias Analysis (FBA), DI FBA, SS Probabilistic Bias Analysis (PBA), DI PBA and MCMC analyses 1 and 2

Method	Bias [minimum,maximum]
Naïve	-0.97
MIME	[-0.28,-0.07]
SS FBA	[-0.72,0.65]
DI FBA	[-0.74,0.77]
SS PBA	[-0.68,0.63]
DI PBA	[-0.53,0.43]
MCMC Analysis 1	[-0.54,0.82]
MCMC Analysis 2	[-0.60,0.39]

b) Probabilistic sensitivity analysis, by including uncertainty in the prior values given to the bias parameters, yields estimates, which are less sensitive to a misspecification of these bias parameters.

In Chapter VI, we observed that in situation 2), when the values for the bias parameters were misspecified in Fixed-parameter Bias analysis (FBA), the adjusted confidence interval often did not include the true OR. On the contrary, Probabilistic Bias Analysis (PBA), by allowing the specification non-informative prior distributions for the bias parameters, yielded credibility intervals, which were much larger and almost always included the true OR.

c) The choice of the method to adjust for misclassification depends on the quantity and quality of information available about the relationship between the 'true' partially or totally missing variable and the other covariates in the study.

Table VII.2 summarizes the strengths and limitations of each method we applied in Chapter VI.

Of course, several other methods or variants of the methods applied here are available to adjust for misclassification, such as regression calibration in situation 1)[203], Predictive Value Weighting in situation 2)[204]. I chose to present and apply these methods here as they are direct extensions of correction methods for missing data and are easy to implement with standard statistical software while accounting for the other covariates present in the study.

Table VII.2. Strengths and limitations of Multiple Imputation for Measurement Error (MIME), Imputation based on Sensitivity and Specificity (SS), Direct Imputation (DI) and Imputation based on a fully Bayesian analysis

Method	Situation	Strengths	Limitations
MIME	1) Validation data available	• No need of *a priori* knowledge	• Cost and time spent for the collection of validation data • The validation subsample needs to be representative of the study population
SS	2) Validation data not available	• No cost and time spent for the collection of validation data • Need for *a priori* knowledge only for the bias parameters (sensitivity and specificity) • Possibility of evaluating the variation of the estimate due to sensitivity and specificity separately	• Need to adjust the *a priori* information so that sensitivity, specificity and prevalences for the 'true' variable are restricted to values between 0 and 1
DI	2) Validation data not available	• No cost and time spent for the collection of validation data • Imputation model coefficients can take any value • Less sensitive than SS to the values taken by the sensitivity and specificity	• Need for *a priori* knowledge for both the bias parameters and the association of interest • Need to translate this *a priori* knowledge into prior values/distributions for the model coefficients • No possibility of evaluating the variation of the estimate due to sensitivity and specificity separately
Fully Bayesian analysis	2) Validation data not available	• No cost and time spent for the collection of validation data • Easy to assign informative/non-informative prior distributions to any parameter	• Complexity of the algorithm

3. Limitations

a) Gender was not a strong confounder in our study

In Chapter VI, the association of interest (odds ratio of lung cancer for ever-smokers vs. never-smokers) was adjusted for gender. Gender was also taken into account in the adjustment for misclassification using MIME, SS, DI and the fully Bayesian analyses. In this study, gender was not a strong confounder of the association between smoking and lung cancer and the inclusion of gender in the adjustment for misclassification did not modify much the odds ratio. Therefore, we cannot predict how the methods would perform in the presence of strong confounding. However, the objective of including gender in the adjustment here was only to illustrate how covariates (categorical or continuous) could be taken into account while adjusting for misclassification by all four methods.

b) Differential misclassification was not considered

In Chapter VI, the smoking status was initially misclassified in a non-differential way. However, the four algorithms could also be easily implemented to adjust for differential misclassification. In MIME and DI, differential misclassification could be addressed by including interaction terms between the misclassified smoking status and the other covariates in the imputation model, and in SS and the fully Bayesian approach, a specification of different prior sensitivities and specificities would have also allowed to adjust for differential misclassification.

D. Future research

Several aspects of the methods presented in this thesis should be investigated further.

a) Further development of the second-stage model in hierarchical regression

In Chapter V, we applied hierarchical regression using a second-stage model that included dichotomous covariates for the level of exposure of each occupation to three lung carcinogens. Shrinkage could be improved by including exposure to carcinogens on a continuous scale in the second-stage model, when this information is made available by the Job Exposure Matrix. Moreover, interactions between first-stage and second-stage covariates such as the interaction between smoking and asbestos should also be considered in the hierarchical model.

b) Correction for the misclassification of confounders

In Chapter VI, we applied several methods to correct for misclassification of the smoking status in the association between smoking and lung cancer, where the smoking status was the exposure of interest. However, in many occupational studies, and in particular studies of lung cancer, such as the study presented in Chapter III, smoking is considered as a confounder in the statistical analyses. Methods described in Chapter VI should also be applied in these studies in order to evaluate the sensitivity of the results to the possible misclassification of the smoking status and should be extended to the case where the smoking status is a three-category variable (never smoker, ex-smoker, current smoker).

E. Conclusions

In summary, there is considerable potential for the use of Bayesian methods to address problems of random error and systematic error, both in occupational cancer epidemiology, and in epidemiology more generally. Most epidemiologists write their methods and results sections as frequentists and their introduction and discussion sections as Bayesians[1]. In their methods and results sections, they "test" their findings as if their data are the only data that exist. In the introduction and discussion, they discuss their findings with regards to their consistency with previous studies, as well as other issues such as biological plausibility. This creates some tensions, e.g. when a small study has findings which are not statistically significant but which are consistent with prior knowledge; or when a study finds statistically significant findings which are inconsistent with prior knowledge. Thus, in practice, almost all epidemiologists profess to be frequentists, but in practice are qualitative Bayesians. In some (but not all) instances, things can be made clearer if we also formally include Bayesian methods in the methods and results sections of our paper, i.e. if we act as quantitative as well as qualitative Bayesians. In this thesis, I have reviewed and applied some of the methods which are currently available, most of which can be easily done with standard statistical analysis packages. Hopefully, in future years, it will become a routine to use EB and SB adjustments for multiple comparisons, and to conduct sensitivity analyses, using standard statistical packages. In most instances, these will not replace our existing methods. Rather they will supplement them, so that we will have an extra paragraph or two in our methods sections, and an extra paragraph or two and additional tables, in our results sections. This will enable us to better quantitatively assess problems of multiple comparisons and bias. These new methods are applicable to all areas of epidemiology, but particularly to occupational epidemiology, because of the large numbers of

comparisons often involved, the hierarchical nature of the exposure data, and the need to quantitatively assess the potential for bias.

References

1. Pearce, N. and M. Corbin, *Why we should be Bayesians (and often already are without realising it).* , in *Current topics in occupational epidemiology.*, K. Venables, Editor. 2013, Oxford University press: Oxford, UK.

2. Greenland, S., *Bayesian perspectives for epidemiological research: I. Foundations and basic methods.* International Journal of Epidemiology, 2006. **35**(3): p. 765-75.

3. Corbin, M., D. McLean, A. Mannetje, et al., *Lung cancer and occupation: A New Zealand cancer registry-based case-control study.* American Journal of Industrial Medicine, 2011. **54**(2): p. 89-101.

4. Richiardi, L., M. Corbin, M. Marron, et al., *Occupation and risk of upper aerodigestive tract cancer: the ARCAGE study.* International Journal of Cancer, 2012. **130**(10): p. 2397-406.

5. Corbin, M., L. Richiardi, R. Vermeulen, et al., *Hierarchical Regression for Multiple Comparisons in a Case-Control Study of Occupational Risks for Lung Cancer.* Plos One, 2012. **7**(6).

6. Corbin, M., S. Haslett, N. Pearce, M. Maule, and S. Greenland, *A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable*. Submitted for publication.

7. Boffetta, P., *Cancer*, in *Encyclopaedia of Occupational Health and Safety: Industries and occupations*. 2011, International Labour Office.

8. Shimizu, H., R.K. Ross, L. Bernstein, R. Yatani, B.E. Henderson, and T.M. Mack, *Cancers of the prostate and breast among Japanese and white immigrants in Los Angeles County*. British journal of cancer, 1991. **63**(6): p. 963-6.

9. Boyle, P. and J.S. Langman, *ABC of colorectal cancer: Epidemiology.* British Medical Journal, 2000. **321**(7264): p. 805-8.

10. Kamineni, A., M.A. Williams, S.M. Schwartz, L.S. Cook, and N.S. Weiss, *The incidence of gastric carcinoma in Asian migrants to the United States and their descendants.* Cancer Causes & Control, 1999. **10**(1): p. 77-83.

11. NCI, *Monograph 1: Strategies to Control Tobacco Use in the United States*, in *Smoking and Tobacco Control Monographs*. 1991, National Cancer Institute: United States.

12. Checkoway, H., N.E. Pearce, and D.L. Kriebel, *Research Methods in Occupational Epidemiology*. 2004: Oxford University Press, Incorporated.

13. IARC, *IARC monographs on the evaluation of carcinogenic risks to humans*. International Agency for Research on Cancer: Lyon, France.

14. Pearce, N. and E. Matos, *Strategies for prevention of occupational cancer in developing countries.*, in *Occupational cancer in developing countries*. 1994, International Agency for Research on Cancer: Lyon, France. p. 173-183.

15. Rothman, K.J., *Epidemiology:An Introduction*. 2002: Oxford University Press, USA.

16. Feise, R.J., *Do multiple outcome measures require p-value adjustment?* BMC Medical Research Methodology, 2002. **2**: p. 8.

17. Corbin, M., M. Maule, L. Richiardi, L. Simonato, F. Merletti, and N. Pearce, *Semi-Bayes and empirical Bayes adjustment methods for multiple comparisons.* Epidemiologia & Prevenzione, 2008. **32**(2): p. 108-110.

18. Woolson, R.F. and W.R. Clarke, *Statistical methods for the analysis of biomedical data*. 2002, Wiley-Interscience. p. 379-80.

19. Perneger, T.V., *What's wrong with Bonferroni adjustments.* British Medical Journal, 1998. **316**(7139): p. 1236-8.

20. Lash, T., M. Fox, and A. Fink, *Applying Quantitative Bias Analysis to Epidemiologic Data*, ed. Springer. 2009.

21. Fisher, R.A., *On the mathematical foundations of theoretical statistics.* Philosophical Transactions of the Royal Society of London. Series A, 1922. **222**(594-604): p. 309-368.

22. Goodman, S.N. and R. Royall, *Evidence and scientific research.* American Journal of Public Health, 1988. **78**(12): p. 1568-74.

23. Lesaffre, E. and A.B. Lawson, *Bayesian Biostatistics*. 2012: Wiley. 534.

24. Bolstad, W.M., *Introduction to Bayesian statistics*. 2nd ed ed. 2007: Hoboken, N.J. : John Wiley, c2007.

25. Christensen, R., W. Johnson, A. Branscum, and T.E. Hanson, *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. 2011: Chapman & Hall /CRC.

26. Kalos, M. and P. Whitlock, *Monte Carlo Methods*. Second ed. 2008, Germany: Wiley-VCH.

27. Marshall, R.J., *Bayesian analysis of case-control studies.* Statistics in Medicine, 1988. **7**(12): p. 1223-30.

28. Zelen, M. and R.A. Parker, *Case-control studies and Bayesian inference.* Statistics in Medicine, 1986. **5**(3): p. 261-9.

29. Ashby, D., J.L. Hutton, and M.A. Mcgee, *Simple Bayesian Analyses for Case-Control Studies in Cancer-Epidemiology.* Statistician, 1993. **42**(4): p. 385-397.

30. Nurminen, M. and P. Mutanen, *Exact Bayesian-Analysis of 2 Proportions.* Scandinavian Journal of Statistics, 1987. **14**(1): p. 67-77.

31. Seaman, S.R. and S. Richardson, *Bayesian analysis of case-control studies with categorical covariates.* Biometrika, 2001. **88**(4): p. 1073-1088.

32. Sinha, S., B. Mukherjee, and M. Ghosh, *Bayesian semiparametric modeling for matched case-control studies with multiple disease states.* Biometrics, 2004. **60**(1): p. 41-9.

33. Mukherjee, B., S. Sinha, and M. Ghosh, *Bayesian analysis of case-control studies.* Handbook of Statistics, 2005. **25**: p. 793-819.

34. Gustafson, P., *Flexible Bayesian modelling for survival data.* Lifetime data analysis, 1998. **4**(3): p. 281-99.

35. Ibrahim, J.G., M.H. Chen, and D. Sinha, *Bayesian Survival Analysis*. 2001: Springer.

36. Carlin, B.P. and J.S. Hodges, *Hierarchical proportional hazards regression models for highly stratified data.* Biometrics, 1999. **55**(4): p. 1162-1170.

37. Greenland, S., *Principles of multilevel modelling.* International Journal of Epidemiology, 2000. **29**(1): p. 158-167.

38. Copas, J.B., *Regression, Prediction and Shrinkage.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1983. **45**(3): p. 311-354.

39. Banerjee, S., B.P. Carlin, and A.E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*. 2003: Chapman & Hall.

40. Lawson, A.B., *Statistical Methods in Spatial Epidemiology*. 2013: Wiley.

41. Wakefield, J., *Disease mapping and spatial regression with count data.* Biostatistics, 2007. **8**(2): p. 158-183.

42. Mollié, A., *Bayesian and empirical Bayes approaches to disease mapping*, in *Disease Mapping and Risk Assessment for Public Health*, A.B. Lawson, et al., Editors. 1999, Wiley: Chichester. p. pp. 15-29.

43. Steenland, K., I. Bray, S. Greenland, and P. Boffetta, *Empirical Bayes adjustments for multiple results in hypothesis-generating or surveillance studies.* Cancer Epidemiology, Biomarkers & Prevention, 2000. **9**(9): p. 895-903.

44. Richiardi, L., P. Boffetta, L. Simonato, et al., *Occupational risk factors for lung cancer in men and women: a population-based case-control study in Italy.* Cancer Causes & Control, 2004. **15**(3): p. 285-94.

45. *International Standard Classification of Occupation. 2nd ed.* 1968, International Labour Organization: Geneva.

46. *International Standard Industrial Classification*. 1971, United Nations: New York: United Nations Statistics Division.

47. Rothman, K.J., *No adjustments are needed for multiple comparisons.* Epidemiology, 1990. **1**(1): p. 43-6.

48. Greenland, S. and C. Poole, *Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance.* Archives of environmental health, 1994. **49**(1): p. 9-16.

49. Chen, G.K. and J.S. Witte, *Enriching the analysis of genomewide association studies with hierarchical modeling.* American Journal of Human Genetics, 2007. **81**(2): p. 397-404.

50. Conti, D.V. and J.S. Witte, *Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations.* American Journal of Human Genetics, 2003. **72**(2): p. 351-63.

51. De Roos, A.J., C. Poole, K. Teschke, and A.F. Olshan, *An application of hierarchical regression in the investigation of multiple paternal occupational exposures and neuroblastoma in offspring.* American Journal of Industrial Medicine, 2001. **39**(5): p. 477-86.

52. Greenland, S., *Hierarchical Regression for Epidemiologic Analyses of Multiple Exposures.* Environmental Health Perspectives, 1994. **102**: p. 33-39.

53. Witte, J.S., S. Greenland, R.W. Haile, and C.L. Bird., *Hierarchical Regression Analysis Applied to a Study of Multiple Dietary Exposures and Breast Cancer.* Epidemiology, 1994. **5**(6): p. 612-21.

54. Witte, J.S. and S. Greenland, *Simulation study of hierarchical regression.* Statistics in Medicine, 1996. **15**(11): p. 1161-70.

55. Witte, J.S., S. Greenland, and L.L. Kim, *Software for hierarchical modeling of epidemiologic data.* Epidemiology, 1998. **9**(5): p. 563-6.

56. Witte, J.S., S. Greenland, L.L. Kim, and L. Arab, *Multilevel modeling in epidemiology with GLIMMIX*. Epidemiology, 2000. **11**(6): p. 684-8.

57. Greenland, S., *Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analyses.* Statistics in Medicine, 1997. **16**(5): p. 515-26.

58. Momoli, F., M. Abrahamowicz, M.E. Parent, D. Krewski, and J. Siemiatycki, *Analysis of multiple exposures: an empirical comparison of results from conventional and semi-bayes modeling strategies.* Epidemiology, 2010. **21**(1): p. 144-51.

59. Dryson, E., A. 't Mannetje, C. Walls, et al., *Case-control study of high risk occupations for bladder cancer in New Zealand.* International Journal of Cancer, 2008. **122**(6): p. 1340-1346.

60. McLean, D., A. 't Mannetje, E. Dryson, et al., *Leukaemia and occupation: a New Zealand Cancer Registry-based casecontrol Study*. International Journal of Epidemiology, 2009. **38**(2): p. 594-606.

61. 't Mannetje, A., E. Dryson, C. Walls, et al., *High risk occupations for non-Hodgkin's lymphoma in New Zealand: case-control study*. Occupational and Environmental Medicine, 2008. **65**(5): p. 354-363.

62. Greenland, S., *Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods.* International Journal of Epidemiology, 2009. **38**(6): p. 1662-1673.

63. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data*. 2nd ed, ed. J.W. Sons. 2002, New York.

64. Allison, P.D., *Missing Data*. 2001: SAGE Publications.

65. Daniels, M.J. and J.W. Hogan, *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. 2008: Taylor & Francis.

66. Molenberghs, G. and M. Kenward, *Missing Data in Clinical Studies*. 2007: Wiley.

67. Tsiatis, A., *Semiparametric Theory and Missing Data*. 2007: Springer.

68. Schafer, J.L., *Analysis of Incomplete Multivariate Data*. 1997: Taylor & Francis.

69. Carpenter, J. and M.U. Kenward, *Multiple Imputation and its Application*. 2012: Wiley.

70. Carpenter, J. and M. Kenward, *Missing data in randomised controlled trials - a practical guide*. 2007.

71. Rubin, D.B., *Multiple imputation for nonresponse in surveys*, ed. J.W. Sons. 1987, New York.

72. Huisman, M., *Missing Data*. 2011.

73. Rubin, D.B., *Multiple imputation after 18+ years.* Journal of the American Statistical Association, 1996. **91**(434): p. 473-489.

74. Cole, S.R., H.T. Chu, and S. Greenland, *Multiple-imputation for measurement-error correction.* International Journal of Epidemiology, 2006. **35**(4): p. 1074-1081.

75. MacLehose, R.F. and P. Gustafson, *Is probabilistic bias analysis approximately Bayesian?* Epidemiology, 2012. **23**(1): p. 151-8.

76. Chu, H., Z. Wang, S.R. Cole, and S. Greenland, *Sensitivity analysis of misclassification: a graphical and a Bayesian approach.* Annals of epidemiology, 2006. **16**(11): p. 834-41.

77. Chu, R., P. Gustafson, and N. Le, *Bayesian adjustment for exposure misclassification in case-control studies.* Statistics in Medicine, 2010. **29**(9): p. 994-1003.

78. Parkin, D.M., F. Bray, J. Ferlay, and P. Pisani, *Global cancer statistics, 2002.* CA: A Cancer Journal for Clinicians, 2005. **55**(2): p. 74-108.

79. Ministry of Health, N.Z., *Cancer in New Zealand -Trends and Projections*. 2002: Wellington.

80. Boffetta, P. and D. Trichopoulos, *Cancer of the Lung, Larynx and Pleura*, in *Textbook of cancer epidemiology*, H. Adami, D. Hunter, and D. Trichopoulos, Editors. 2002, Oxford University Press: New York. p. p248-280.

81. Spitz, M.R., X. Wu, A. Wilkinson, and Q. Wei, *Cancer of the lung*, in *Cancer Epidemiology and Prevention (3rd Ed.)*, D. Schottenfeld and J.F. Fraumeni, Editors. 2006, Oxford University Press: New York.

82. McLean, D. and N. Pearce, *Cancer among meat industry workers.* Scandinavian Journal of Work, Environment & Health, 2004. **30**(6): p. 425-37.

83. Kawachi, I., N. Pearce, and J. Fraser, *A New Zealand Cancer Registry-based study of cancer in wood workers.* Cancer, 1989. **64**(12): p. 2609-13.

84. Statistics New Zealand, *New Zealand Standard Classification of Occupations 1999*. 2001: Wellington.

85. Statistics New Zealand, *Australian and New Zealand Standard Industrial Classification (New Zealand Use Version) 1996. Version 4.1*. 2004: Wellington.

86. Steenland, K., C. Burnett, N. Lalich, E. Ward, and J. Hurrell, *Dying for work: The magnitude of US mortality from selected causes of death associated with occupation.* American Journal of Industrial Medicine, 2003. **43**(5): p. 461-82.

87. Nurminen, M. and A. Karjalainen, *Epidemiologic estimate of the proportion of fatalities related to occupational factors in Finland.* Scandinavian Journal of Work, Environment & Health, 2001. **27**(3): p. 161-213.

88. 't Mannetje, A. and N. Pearce, *Quantitative estimates of work-related death, disease and injury in New Zealand.* Scandinavian Journal of Work, Environment & Health, 2005. **31**(4): p. 266-76.

89. Davis, P., K. McLeod, M. Ransom, P. Ongley, N. Pearce, and P. Howden-Chapman, *The New Zealand Socioeconomic Index: developing and validating an occupationally-derived indicator of socio-economic status.* Australian and New Zealand Journal of Public Health, 1999. **23**(1): p. 27-33.

90. R Foundation for Statistical Computing, *R: A language and environment for statistical computing*. 2006: Vienna, Australia.

91. 't Mannetje, A., D. McLean, and N. Pearce, *The development of a job-exposure-matrix for New Zealand.* Arbo Wetenschap, 2004. **suppl. 2; 38 (abstract)**.

92. Ministry of Social Development, N.Z., *Unemployment*, in *2009 the social report*. 2009.

93. Richiardi, L., F. Forastiere, P. Boffetta, L. Simonato, and F. Merletti, *Effect of different approaches to treatment of smoking as a potential confounder in a case-control study on occupational exposures.* Occupational and Environmental Medicine, 2005. **62**(2): p. 101-4.

94. IARC, *IARC monographs on the evaluation of carcinogenic risk to humans. Wood dust and formaldehyde.* 1995, International Agency for Research on Cancer: Lyon, France.

95. Jahn, I., W. Ahrens, I. Bruske-Hohlfeld, et al., *Occupational risk factors for lung cancer in women: results of a case-control study in Germany.* American Journal of Industrial Medicine, 1999. **36**(1): p. 90-100.

96. Dement, J., L. Pompeii, I.M. Lipkus, and G.P. Samsa, *Cancer incidence among union carpenters in New Jersey.* Journal of Occupational and Environmental Medicine, 2003. **45**(10): p. 1059-67.

97. Barcenas, C.H., G.L. Delclos, R. El-Zein, G. Tortolero-Luna, L.W. Whitehead, and M.R. Spitz, *Wood dust exposure and the association with lung cancer risk.* American Journal of Industrial Medicine, 2005. **47**(4): p. 349-57.

98. Morabia, A., S. Markowitz, K. Garibaldi, and E.L. Wynder, *Lung cancer and occupation: results of a multicentre case-control study.* British Journal of Industrial Medicine, 1992. **49**(10): p. 721-7.

99. Bruske-Hohlfeld, I., M. Mohner, H. Pohlabeln, et al., *Occupational lung cancer risk for men in Germany: results from a pooled case-control study.* American Journal of Epidemiology, 2000. **151**(4): p. 384-95.

100. Matos, E.L., M. Vilensky, D. Mirabelli, and P. Boffetta, *Occupational exposures and lung cancer in Buenos Aires, Argentina.* Journal of Occupational and Environmental Medicine, 2000. **42**(6): p. 653-9.

101. Bardin-Mikolajczak, A., J. Lissowska, D. Zaridze, et al., *Occupation and risk of lung cancer in Central and Eastern Europe: the IARC multi-center case-control study.* Cancer Causes & Control, 2007. **18**(6): p. 645-54.

102. MacArthur, A.C., N.D. Le, R. Fang, and P.R. Band, *Identification of occupational cancer risk in British Columbia: a population-based case-control study of 2,998 lung cancers by histopathological subtype.* American Journal of Industrial Medicine, 2009. **52**(3): p. 221-32.

103. Yenugadhati, N., N.J. Birkett, F. Momoli, and D. Krewski, *Occupations and lung cancer: a population-based case-control study in British Columbia.* Journal of Toxicology and Environmental Health, Part A, 2009. **72**(10): p. 658-75.

104. IARC, *IARC monographs on the evaluation of carcinogenic risks to humans. Chromium, nickel and welding.* 1990, International Agency for Research on Cancer: Lyon, France.

105. Siemiatycki, J., L. Richardson, K. Straif, et al., *Listing occupational carcinogens.* Environmental Health Perspectives, 2004. **112**(15): p. 1447-59.

106. Benhamou, S., E. Benhamou, and R. Flamant, *Occupational risk factors of lung cancer in a French case-control study.* British Journal of Industrial Medicine, 1988. **45**(4): p. 231-3.

107. Moulin, J.J., *A meta-analysis of epidemiologic studies of lung cancer in welders.* Scandinavian Journal of Work, Environment & Health, 1997. **23**(2): p. 104-13.

108. Jockel, K.H., W. Ahrens, H. Pohlabeln, U. Bolm-Audorff, and K.M. Muller, *Lung cancer risk and welding: results from a case-control study in Germany.* American Journal of Industrial Medicine, 1998. **33**(4): p. 313-20.

109. Becker, N., *Cancer mortality among arc welders exposed to fumes containing chromium and nickel. Results of a third follow-up: 1989-1995.* Journal of Occupational and Environmental Medicine, 1999. **41**(4): p. 294-303.

110. Coggon, D., B. Pannett, E.C. Pippard, and P.D. Winter, *Lung cancer in the meat industry.* British Journal of Industrial Medicine, 1989. **46**(3): p. 188-91.

111. Fritschi, L., S. Fenwick, and M. Bulsara, *Mortality and cancer incidence in a cohort of meatworkers.* Occupational and Environmental Medicine, 2003. **60**(9): p. E4.

112. Durusoy, R., P. Boffetta, A. 't Mannetje, et al., *Lung cancer risk and occupational exposure to meat and live animals.* International Journal of Cancer, 2006. **118**(10): p. 2543-7.

113. McLean, D., S. Cheng, A. 't Mannetje, A. Woodward, and N. Pearce, *Mortality and cancer incidence in New Zealand meat workers.* Occupational and Environmental Medicine, 2004. **61**(6): p. 541-7.

114. Travier, N., G. Gridley, A.J. De Roos, N. Plato, T. Moradi, and P. Boffetta, *Cancer incidence of dry cleaning, laundry and ironing workers in Sweden.* Scandinavian Journal of Work, Environment & Health, 2002. **28**(5): p. 341-8.

115. Blair, A., P. Decoufle, and D. Grauman, *Causes of death among laundry and dry cleaning workers.* American Journal of Public Health, 1979. **69**(5): p. 508-11.

116. Duh, R.W. and N.R. Asal, *Mortality among laundry and dry cleaning workers in Oklahoma.* American Journal of Public Health, 1984. **74**(11): p. 1278-80.

117. Ruder, A.M., E.M. Ward, and D.P. Brown, *Mortality in dry-cleaning workers: an update.* American Journal of Industrial Medicine, 2001. **39**(2): p. 121-32.

118. IARC, *IARC monographs on the evaluation of the carcinogenic risks to humans. Dry cleaning, some chlorinated solvents and other industrial chemicals.* 1995: Lyon, France.

119. Mastrangelo, G., U. Fedeli, E. Fadda, G. Milan, and J.H. Lange, *Epidemiologic evidence of cancer risk in textile industry workers: a review and update.* Toxicology and Industrial Health, 2002. **18**(4): p. 171-81.

120. Mastrangelo, G., E. Fadda, R. Rylander, et al., *Lung and other cancer site mortality in a cohort of Italian cotton mill workers.* Occupational and Environmental Medicine, 2008. **65**(10): p. 697-700.

121. Su, W.L., Y.H. Chen, S.H. Liou, and C.P. Wu, *Meta-analysis of standard mortality ratio in cotton textile workers.* European Journal of Epidemiology, 2004. **19**(11): p. 989-97.

122. Astrakianakis, G., N.S. Seixas, R. Ray, et al., *Lung cancer risk among female textile workers exposed to endotoxin.* Journal of the National Cancer Institute, 2007. **99**(5): p. 357-64.

123. Kuzmickiene, I. and M. Stukonis, *Lung cancer risk among textile workers in Lithuania.* Journal of Occupational Medicine and Toxicology, 2007. **2**: p. 14.

124. Dement, J.M., D.P. Brown, and A. Okun, *Follow-up study of chrysotile asbestos textile workers: cohort mortality and case-control analyses.* American Journal of Industrial Medicine, 1994. **26**(4): p. 431-47.

125. Hein, M.J., L.T. Stayner, E. Lehman, and J.M. Dement, *Follow-up study of chrysotile textile workers: cohort mortality and exposure-response.* Occupational and Environmental Medicine, 2007. **64**(9): p. 616-25.

126. Loomis, D., J.M. Dement, S.H. Wolf, and D.B. Richardson, *Lung Cancer Mortality and Fiber Exposures among North Carolina Asbestos Textile Workers.* Occupational and Environmental Medicine, 2009.

127. IARC, *IARC monographs on the evaluation of carcinogenic risks to humans. Polynuclear aromatic compounds, engine exhausts and nitroarenes.* 1989, International Agency for Research on Cancer: Lyon, France.

128. Hayes, R.B., T. Thomas, D.T. Silverman, et al., *Lung cancer in motor exhaust-related occupations.* American Journal of Industrial Medicine, 1989. **16**(6): p. 685-95.

129. Steenland, N.K., D.T. Silverman, and R.W. Hornung, *Case-control study of lung cancer and truck driving in the Teamsters Union.* American Journal of Public Health, 1990. **80**(6): p. 670-4.

130. Finkelstein, M.M., *Occupational associations with lung cancer in two Ontario cities.* American Journal of Industrial Medicine, 1995. **27**(1): p. 127-36.

131. Hansen, J., O. Raaschou-Nielsen, and J.H. Olsen, *Increased risk of lung cancer among different types of professional drivers in Denmark.* Occupational and Environmental Medicine, 1998. **55**(2): p. 115-8.

132. Menvielle, G., D. Luce, J. Fevotte, et al., *Occupational exposures and lung cancer in New Caledonia.* Occupational and Environmental Medicine, 2003. **60**(8): p. 584-9.

133. Doebbert, G., K.R. Riedmiller, and K.W. Kizer, *Occupational mortality of California women, 1979-1981.* Western Journal of Medicine, 1988. **149**(6): p. 734-40.

134. Petralia, S.A., M. Dosemeci, E.E. Adams, and S.H. Zahm, *Cancer mortality among women employed in health care occupations in 24 U.S. states, 1984-1993.* American Journal of Industrial Medicine, 1999. **36**(1): p. 159-65.

135. Milne, K.L., D.P. Sandler, R.B. Everson, and S.M. Brown, *Lung cancer and occupation in Alameda County: a death certificate case-control study.* American Journal of Industrial Medicine, 1983. **4**(4): p. 565-75.

136. Sun, J., H. Kubota, N. Hisanaga, E. Shibata, M. Kamijima, and K. Nakamura, *Mortality among Japanese construction workers in Mie Prefecture.* Occupational and Environmental Medicine, 2002. **59**(8): p. 512-6.

137. IARC, *IARC monograph on the evaluation of the carcinogenic risk of chemicals to humans: alcohol consumption and ethyl carbamate (Urethane)*. 2010, International Agency for Research on Cancer: Lyon, France.

138. IARC, *IARC Monograph on the evaluation of the carcinogenic risk of chemicals to humans: tobacco smoke and involuntary smoking*. 2004, International Agency for Research on Cancer: Lyon, France.

139. Hashibe, M., P. Brennan, S.C. Chuang, et al., *Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium.* Cancer Epidemiology, Biomarkers & Prevention, 2009. **18**(2): p. 541-50.

140. Lagiou, P., R. Talamini, E. Samoli, et al., *Diet and upper-aerodigestive tract cancer in Europe: the ARCAGE study.* International Journal of Cancer, 2009. **124**(11): p. 2671-6.

141. IARC, *IARC monograph on the evaluation of the carcinogenic risk of chemicals to humans: human papillomaviruses*. 2007, International Agency for Research on Cancer: Lyon, France.

142. Conway, D.I., P.A. McKinney, A.D. McMahon, et al., *Socioeconomic factors associated with risk of upper aerodigestive tract cancer in Europe.* European journal of cancer, 2010. **46**(3): p. 588-98.

143. Canova, C., M. Hashibe, L. Simonato, et al., *Genetic associations of 115 polymorphisms with cancers of the upper aerodigestive tract across 10 European countries: the ARCAGE project.* Cancer research, 2009. **69**(7): p. 2956-65.

144. Elci, O.C., M. Akpinar-Elci, A. Blair, and M. Dosemeci, *Occupational dust exposure and the risk of laryngeal cancer in Turkey.* Scandinavian Journal of Work, Environment & Health, 2002. **28**(4): p. 278-84.

145. Wortley, P., T.L. Vaughan, S. Davis, M.S. Morgan, and D.B. Thomas, *A case-control study of occupational risk factors for laryngeal cancer.* British Journal of Industrial Medicine, 1992. **49**(12): p. 837-44.

146. Zagraniski, R.T., J.L. Kelsey, and S.D. Walter, *Occupational risk factors for laryngeal carcinoma: Connecticut, 1975-1980.* American Journal of Epidemiology, 1986. **124**(1): p. 67-76.

147. Becher, H., H. Ramroth, W. Ahrens, A. Risch, P. Schmezer, and A. Dietz, *Occupation, exposure to polycyclic aromatic hydrocarbons and laryngeal cancer risk.* International Journal of Cancer, 2005. **116**(3): p. 451-7.

148. Brown, L.M., T.J. Mason, L.W. Pickle, et al., *Occupational risk factors for laryngeal cancer on the Texas Gulf Coast.* Cancer research, 1988. **48**(7): p. 1960-4.

149. Goldberg, P., A. Leclerc, D. Luce, J.F. Morcet, and J. Brugere, *Laryngeal and hypopharyngeal cancer and occupation: results of a case control-study.* Occupational and Environmental Medicine, 1997. **54**(7): p. 477-82.

150. Boffetta, P., L. Richiardi, F. Berrino, et al., *Occupation and larynx and hypopharynx cancer: an international case-control study in France, Italy, Spain, and Switzerland.* Cancer Causes & Control, 2003. **14**(3): p. 203-12.

151. Flanders, W.D., C.I. Cann, K.J. Rothman, and M.P. Fried, *Work-related risk factors for laryngeal cancer.* American Journal of Epidemiology, 1984. **119**(1): p. 23-32.

152. Flanders, W.D. and K.J. Rothman, *Occupational risk for laryngeal cancer.* American Journal of Public Health, 1982. **72**(4): p. 369-72.

153. Haguenoer, J.M., S. Cordier, C. Morel, J.L. Lefebvre, and D. Hemon, *Occupational risk factors for upper respiratory tract and upper digestive tract cancers.* British Journal of Industrial Medicine, 1990. **47**(6): p. 380-3.

154. Muscat, J.E. and E.L. Wynder, *Tobacco, alcohol, asbestos, and occupational risk factors for laryngeal cancer.* Cancer, 1992. **69**(9): p. 2244-51.

155. Ahrens, W., K.H. Jockel, W. Patzak, and G. Elsner, *Alcohol, smoking, and occupational factors in cancer of the larynx: a case-control study.* American Journal of Industrial Medicine, 1991. **20**(4): p. 477-93.

156. Olsen, J. and S. Sabroe, *Occupational causes of laryngeal cancer.* Journal of epidemiology and community health, 1984. **38**(2): p. 117-21.

157. De Stefani, E., P. Boffetta, F. Oreggia, A. Ronco, M. Kogevinas, and M. Mendilaharsu, *Occupation and the risk of laryngeal cancer in Uruguay.* American Journal of Industrial Medicine, 1998. **33**(6): p. 537-42.

158. Huebner, W.W., J.B. Schoenberg, J.L. Kelsey, et al., *Oral and pharyngeal cancer and occupation: a case-control study.* Epidemiology, 1992. **3**(4): p. 300-9.

159. Zheng, W., W.J. Blot, X.O. Shu, et al., *Diet and other risk factors for laryngeal cancer in Shanghai, China.* American Journal of Epidemiology, 1992. **136**(2): p. 178-91.

160. Elci, O.C., M. Dosemeci, and A. Blair, *Occupation and the risk of laryngeal cancer in Turkey.* Scandinavian Journal of Work, Environment & Health, 2001. **27**(4): p. 233-9.

161. Pukkala, E., J.I. Martinsen, E. Lynge, et al., *Occupation and cancer - follow-up of 15 million people in five Nordic countries.* Acta oncologica, 2009. **48**(5): p. 646-790.

162. Berrino, F., L. Richiardi, P. Boffetta, et al., *Occupation and larynx and hypopharynx cancer: a job-exposure matrix approach in an international case-control study in France, Italy, Spain and Switzerland.* Cancer Causes & Control, 2003. **14**(3): p. 213-23.

163. Shangina, O., P. Brennan, N. Szeszenia-Dabrowska, et al., *Occupational exposure and laryngeal and hypopharyngeal cancer risk in central and eastern Europe.* American Journal of Epidemiology, 2006. **164**(4): p. 367-75.

164. Lagiou, P., C. Georgila, P. Minaki, et al., *Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection.* European Journal of Cancer Prevention, 2009. **18**(1): p. 76-84.

165. *National industrial classification of all economic activities (NACE), rev. 1, 2nd ed.* 1993, European Commission, Office for Official Publications of the EC: Luxembourg.

166. Richiardi, L., F. Barone-Adesi, F. Merletti, and N. Pearce, *Using directed acyclic graphs to consider adjustment for socioeconomic status in occupational cancer studies.* Journal of epidemiology and community health, 2008. **62**(7): p. e14.

167. Boffetta, P., I. Burstyn, T. Partanen, et al., *Cancer mortality among European asphalt workers: an international epidemiological study. I. Results of the analysis based on job titles.* American Journal of Industrial Medicine, 2003. **43**(1): p. 18-27.

168. Behrens, T., W. Schill, and W. Ahrens, *Elevated cancer mortality in a German cohort of bitumen workers: extended follow-up through 2004.* Journal of occupational and environmental hygiene, 2009. **6**(9): p. 555-61.

169. Dietz, A., H. Ramroth, T. Urban, W. Ahrens, and H. Becher, *Exposure to cement dust, related occupational groups and laryngeal cancer risk: results of a population based case-control study.* International Journal of Cancer, 2004. **108**(6): p. 907-11.

170. Purdue, M.P., B. Jarvholm, I.A. Bergdahl, R.B. Hayes, and D. Baris, *Occupational exposures and head and neck cancers among Swedish construction workers.* Scandinavian Journal of Work, Environment & Health, 2006. **32**(4): p. 270-5.

171. Steenland, K. and S. Palu, *Cohort mortality study of 57,000 painters and other union members: a 15 year update.* Occupational and Environmental Medicine, 1999. **56**(5): p. 315-21.

172. IARC, *IARC monograph on the evaluation of the carcinogenic risk of chemicals to humans: painting, firefighting, and shiftwork*. 2010, International Agency for Research on Cancer: Lyon, France.

173. Guberan, E., M. Usel, L. Raymond, R. Tissot, and P.M. Sweetnam, *Disability, mortality, and incidence of cancer among Geneva painters and electricians: a historical prospective study.* British Journal of Industrial Medicine, 1989. **46**(1): p. 16-23.

174. Darby, S.C., E. Whitley, G.R. Howe, et al., *Radon and cancers other than lung cancer in underground miners: a collaborative analysis of 11 studies.* Journal of the National Cancer Institute, 1995. **87**(5): p. 378-84.

175. Kreuzer, M., L. Walsh, M. Schnelzer, A. Tschense, and B. Grosche, *Radon and risk of extrapulmonary cancers: results of the German uranium miners' cohort study, 1960-2003.* British journal of cancer, 2008. **99**(11): p. 1946-53.

176. Vacquier, B., S. Caer, A. Rogel, et al., *Mortality risk in the French cohort of uranium miners: extended follow-up 1946-1999.* Occupational and Environmental Medicine, 2008. **65**(9): p. 597-604.

177. Mohner, M., M. Lindtner, and H. Otten, *Ionizing radiation and risk of laryngeal cancer among German uranium miners.* Health physics, 2008. **95**(6): p. 725-33.

178. Laakkonen, A., T. Kauppinen, and E. Pukkala, *Cancer risk among Finnish food industry workers.* International Journal of Cancer, 2006. **118**(10): p. 2567-71.

179. Reif, J., N. Pearce, and J. Fraser, *Cancer risks in New Zealand farmers.* International Journal of Epidemiology, 1989. **18**(4): p. 768-74.

180. Stark, A.D., H.G. Chang, E.F. Fitzgerald, K. Riccardi, and R.R. Stone, *A retrospective cohort study of cancer incidence among New York State Farm Bureau members.* Archives of environmental health, 1990. **45**(3): p. 155-62.

181. Rafnsson, V. and H. Gunnarsdottir, *Mortality among farmers in Iceland.* International Journal of Epidemiology, 1989. **18**(1): p. 146-51.

182. Greenland, S. and J.M. Robins, *Empirical-Bayes adjustments for multiple comparisons are sometimes useful.* Epidemiology, 1991. **2**(4): p. 244-51.

183. Greenland, S., *A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study.* Statistics in Medicine, 1992. **11**(2): p. 219-30.

184. IARC, *IARC monographs - Classifications*. 2011, International Agency for Research on Cancer (IARC): Lyon, France.

185. Peters, S., R. Vermeulen, A. Cassidy, et al., *Comparison of exposure assessment methods for occupational carcinogens in a multi-centre lung cancer case-control study.* Occupational and Environmental Medicine, 2011. **68**(2): p. 148-53.

186. Greenland, S., *Bayesian perspectives for epidemiological research. II. Regression analysis.* International Journal of Epidemiology, 2007. **36**(1): p. 195-202.

187. Guha, N., F. Merletti, N.K. Steenland, A. Altieri, V. Cogliano, and K. Straif, *Lung cancer risk in painters: a meta-analysis.* Environmental Health Perspectives. **118**(3): p. 303-12.

188. Rothman, K.J., S. Greenland, and T.L. Lash, *Validity in Epidemiologic Studies*, in *Modern Epidemiology*. 2008, Wolters Kluwer Health/Lippincott Williams & Wilkins.

189. White, I.R., P. Royston, and A.M. Wood, *Multiple imputation using chained equations: Issues and guidance for practice.* Statistics in Medicine, 2011. **30**(4): p. 377-99.

190. Carroll, R.J., D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. 2 edition ed. 2006, Boca Raton: Chapman and Hall/CRC.

191. Greenland, S. and T.L. Lash, *Bias Analysis*, in *Modern Epidemiology*, K.J. Rothman, S. Greenland, and T.L. Lash, Editors. 2008, Wolters Kluwer Health/Lippincott Williams & Wilkins.

192. Bross, I., *Misclassification in 2 X 2 Tables.* Biometrics, 1954. **10**(4): p. 478-486.

193. Marshall, R.J., *Validation-Study Methods for Estimating Exposure Proportions and Odds Ratios with Misclassified Data.* Journal of clinical epidemiology, 1990. **43**(9): p. 941-947.

194. Kleinbaum, D.G., L.L. Kupper, and H. Morgenstern, *Epidemiologic Research: Principles and Quantitative Methods*. 1982: Wiley.

195. Fox, M.P., T.L. Lash, and S. Greenland, *A method to automate probabilistic sensitivity analyses of misclassified binary variables.* International Journal of Epidemiology, 2005. **34**(6): p. 1370-6.

196. Greenland, S., *Multiple-bias modelling for analysis of observational data.* Journal of the Royal Statistical Society Series a-Statistics in Society, 2005. **168**: p. 267-291.

197. Patrick, D.L., A. Cheadle, D.C. Thompson, P. Diehr, T. Koepsell, and S. Kinne, *The validity of self-reported smoking: a review and meta-analysis.* American Journal of Public Health, 1994. **84**(7): p. 1086-93.

198. Blettner, M. and J. Wahrendorf, *What does an observed relative risk convey about possible misclassification?* Methods of information in medicine, 1984. **23**(1): p. 37-40.

199. Greenland, S. and L. Kheifets, *Leukemia attributable to residential magnetic fields: results from analyses allowing for study biases.* Risk Analysis, 2006. **26**(2): p. 471-82.

200. Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap*. 1994, New York: Chapman and Hall/CRC.

201. Greenland, S., *Prior data for non-normal priors.* Statistics in Medicine, 2007. **26**(19): p. 3578-90.

202. Lash, T.L., M.P. Fox, R.F. MacLehose, G. Maldonado, L.C. McCandless, and S. Greenland, *Good practices for quantitative bias analysis.* International Journal of Epidemiology, 2014. **43**.

203. Rosner, B., W.C. Willett, and D. Spiegelman, *Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error.* Statistics in Medicine, 1989. **8**(9): p. 1051-69; discussion 1071-3.

204. Lyles, R.H. and J. Lin, *Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting.* Statistics in Medicine, 2010. **29**(22): p. 2297-309.

Appendices

List of Appendices

Appendix I – Publications arising from the work presented in the thesis

Corbin M, Maule M, Richiardi L, Simonato L, Merletti F, Pearce N. Semi-Bayes adjustment methods for multiple comparisons. Epidemiologia e Prevenzione 2008; 32: 108-110.

Corbin M, McLean D, 't Mannetje A, Dryson E, Walls C, McKenzie F, Maule M, Cheng S, Cunningham C, Kromhout H, Blair A, Pearce N. Lung cancer and occupation: a New Zealand Cancer Registry-based case-control study. American Journal of Industrial Medicine 2011; 54: 89-101.

Richiardi, L., Corbin, M., Marron, M., Ahrens, W., Pohlabeln, H., Lagiou, P., . . . Merletti, F. (2012). Occupation and risk of upper aerodigestive tract cancer: the ARCAGE study. International Journal of Cancer, 130(10), 2397-2406. doi: 10.1002/ijc.26237

Corbin M, Richiardi L, Vermeulen R, Kromhout H, Merletti F, Peters S, Simonato L, Steenland K, Pearce N, Maule M. Hierarchical regression for multiple comparisons in a case-control study of occupational risks for lung cancer. PLoS One 2012; 7: e38944. doi: 10.1371/journal.pone.0038944.

Pearce N, Corbin M. Why we should be Bayesians (and often already are without realising it). In: Venables K (ed). Current topics in occupational epidemiology. Oxford: Oxford University press, 2013, pp 218-233.

Corbin M, Haslett S, Pearce N, Maule M, Greenland S. A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable
– *Submitted for publication* –

Appendix II – Further details of methodology

Appendix AII.1. Supporting information for Chapter V

1. Section of the matrix $\mathbf{Z}$ for six occupations (rows 55 to 60)

Occupation	Z's row	Elements of $\mathbf{Z}$						
		Intercept	Asbestos *low* (1)	Asbestos *high* (2)	Chromium *low* (1)	Chromium *high* (2)	Silica *low* (1)	Silica *high* (2)
627-nursery workers and gardeners	55	1	0	0	0	0	1	0
628-farm machinery operators	56	1	0	0	0	0	0	1
631-loggers	57	1	0	0	0	0	0	0
641-fishermen	58	1	0	0	0	0	0	0
700-production supervisors and general foremen	59	1	0	0	0	0	0	0
711-miners and quarrymen	60	1	1	0	0	0	0	1

2. Examples of calculation of the elements of the second-stage covariance matrix

$\tau^2\mathbf{T}$

Occupation	Z's: i^{th} row	Carcinogenic exposure			t_{ii}	Second-stage residual variance $\tau^2 t_{ii}$			
		ASB*	CR*	SI*		τ =0.76	τ =0.59	τ =0.41	τ =0.23
627-nursery workers and gardeners	55	0	0	1	0.72	0.42	0.25	0.12	0.04
628-farm machinery operators	56	0	0	2	0.57	0.33	0.20	0.10	0.03
631-loggers	57	0	0	0	1	0.58	0.35	0.17	0.05
641-fishermen	58	0	0	0	1	0.58	0.35	0.17	0.05
700-production supervisors and general foremen	59	0	0	0	1	0.58	0.35	0.17	0.05
711-miners and quarrymen	60	1	0	2	0.37	0.21	0.13	0.06	0.02

*ASB=Asbestos(0 = no exposure, 1 = low exposure, 2 = high exposure)
CR=Chromium(0 = no exposure, 1 = low exposure, 2 = high exposure)
SI=Silica(0 = no exposure, 1 = low exposure, 2 = high exposure)

3. Computation of the Hierarchical Regression estimates

The second-stage coefficients π for the different categories of exposures to the three carcinogens are estimated through weighted least squares with

$$\tilde{\pi} = \left(\mathbf{Z}'\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\hat{\beta} \qquad \text{(A1)}$$

where $\mathbf{W} = \left[\hat{\mathbf{V}} + \tau^2\mathbf{T}\right]^{-1}$ and $\hat{\mathbf{V}}$ is a diagonal matrix composed of the estimated variances of the first stage coefficients estimated in the conventional analysis $\hat{\beta}$.

The Hierarchical Regression estimates $\tilde{\beta}$ are then obtained by averaging the first stage coefficients $\hat{\beta}$ with their respective prior means $\mathbf{Z}\tilde{\pi}$

$$\tilde{\beta} = \mathbf{B}\mathbf{Z}\tilde{\pi} + \left(\mathbf{I} - \mathbf{B}\right)\hat{\beta}, \text{ where } \mathbf{B} = \mathbf{W}\hat{\mathbf{V}} = \left(\hat{\mathbf{V}} + \tau^2\mathbf{T}\right)^{-1}\hat{\mathbf{V}}. \qquad \text{(A2)}$$

Their covariance matrix is estimated by $\tilde{\mathbf{C}} = \hat{\mathbf{V}}\left(\mathbf{I} - \left(\mathbf{I} - \mathbf{H}\right)'\mathbf{B}\right)$ where

$$\mathbf{H} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}$$

Example of calculation of the Hierarchical Regression estimates for the miners and quarrymen (ISCO: 711) (60th row of $\mathbf{Z}$)

		Maximum-likelihood			Hierarchical regression	
	$\mathbf{Z}_{60} \times \tilde{\pi}$	$\hat{\beta}_{60}$	OR_{60}	b_{ii}*	$\tilde{\beta}_{60}$	OR_{60}
τ =0.76	0.31	0.17	1.19	0.39	0.22	1.25
τ =0.59	0.31	0.17	1.19	0.52	0.24	1.27
τ =0.41	0.31	0.17	1.19	0.70	0.27	1.30
τ =0.23	0.31	0.17	1.19	0.87	0.29	1.34

*diagonal elements of the matrix

Appendix AII.2. Supporting information for Chapter VI

1. Details of the calculation of the values for model (2) coefficients in DI FBA

As mentioned in Table VI.4, we assume nondifferentiality i.e. $se_{00}=se_{01}=se_{10}=se_{11}=se$ and $sp_{00}=sp_{01}=sp_{10}=sp_{11}=sp$, where se_{yc} and sp_{yc} are the sensitivity and the specificity for $Y=y$ and $C=c$.

According to the 2009 New Zealand Tobacco Use Survey (NZTUS), the prevalence $P(T=1|Y=0,C=0)$ of ever-smokers (current smokers and ex-smokers) among women in New Zealand is 0.403 and the prevalence $P(T=1|Y=0,C=1)$ of ever-smokers among men in New Zealand is 0.428.

Model (2) coefficients β_0, β_X and β_Y are assigned to different sets of values while coefficients β_Y and β_{YC} are set constant.

Calculation of β_0:

$$\beta_0 = \log\left(\frac{(1-se)\times P(T=1|Y=0,C=0)}{sp\times(1-P(T=1|Y=0,C=0))}\right)$$ (see Table VI.4)

- Sets of values 1,4,7: $\beta_0 = \log\left(\frac{0.3\times 0.403}{0.8\times 0.597}\right) = -1.37$
- Sets of values 2,5,8: $\beta_0 = \log\left(\frac{0.2\times 0.403}{0.9\times 0.597}\right) = -1.90$
- Sets of values 3,6,9: $\beta_0 = \log\left(\frac{0.1\times 0.403}{0.99\times 0.597}\right) = -2.69$

Calculation of β_X :

$\beta_X = \log\left(\frac{se \times sp}{(1-se)\times(1-sp)}\right)$ (see Table VI.4)

- Sets of values 1,4,7: $\beta_X = \log\left(\frac{0.7\times 0.8}{0.3\times 0.2}\right) \approx 2.23$
- Sets of values 2,5,8: $\beta_X = \log\left(\frac{0.8\times 0.9}{0.2\times 0.1}\right) \approx 3.58$
- Set of values 3,6,9: $\beta_X = \log\left(\frac{0.9\times 0.99}{0.1\times 0.01}\right) \approx 6.79$

Calculation of β_Y :

$\beta_Y = \log\left(OR_{TY}(C=0)\right)$ (see Table VI.4)

- Sets of values 1,2,3: $\beta_Y = \log(6.93) \approx 1.94$ (where 6.93 is the smoking-lung cancer odds ratio for women in our original data)
- Set of values 4,5,6: $\beta_Y = \log(3.5) \approx 1.25$
- Set of values 7,8,9: $\beta_Y = \log(14) \approx 2.64$

Calculation of β_C :

$$\beta_C = \log\left(OR_{TC}(Y=0)\right) = \frac{\frac{P(T=1|C=1,Y=0)}{1-P(T=1|C=1,Y=0)}}{\frac{P(T=1|C=0,Y=0)}{1-P(T=1|C=0,Y=0)}} = \frac{\frac{0.428}{0.572}}{\frac{0.403}{0.597}} \approx 0.10$$

Calculation of β_{YC}:

$\beta_{YC} = \log\left(\frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}\right) = \frac{11.03}{6.93} \approx 0.46$ (where 11.03 and 6.93 are the smoking-lung cancer odds ratios for men and women, respectively, in our original data)

2. Details of the calculation of the prior distribution parameters for model (2) coefficients in DI PBA

Let:

- $\lambda^{mean}, \lambda^{sd}, \varepsilon^{mean}, \varepsilon^{sd}$ be the prior means and standard deviations given to MCMC analysis 2 parameters λ and ε, respectively.
- ${\alpha_0}^{mean}$ and ${\alpha_0}^{sd}$ be the prior mean and standard deviation given to MCMC analysis 2 parameter α_0.
- $OR_{TY}(C=0)^{mean}, OR_{TC}(Y=0)^{mean}, \frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}^{mean}$ be the means given to $\mathrm{OR}_{TY}(C=0),\ \mathrm{OR}_{TC}(Y=0),\ \frac{\mathrm{OR}_{TY}(C=1)}{\mathrm{OR}_{TY}(C=0)}$, respectively (see Table VI.5).
- $OR_{TY}(C=0)^{lower}, OR_{TC}(Y=0)^{lower}, \frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}^{lower}$ be the lower 95% limits given to $\mathrm{OR}_{TY}(C=0),\ \mathrm{OR}_{TC}(Y=0),\ \frac{\mathrm{OR}_{TY}(C=1)}{\mathrm{OR}_{TY}(C=0)}$, respectively (see Table VI.5).
- $OR_{TY}(C=0)^{upper}, OR_{TC}(Y=0)^{upper}, \frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}^{upper}$ be the upper 95% limits given to $\mathrm{OR}_{TY}(C=0),\ \mathrm{OR}_{TC}(Y=0),\ \frac{\mathrm{OR}_{TY}(C=1)}{\mathrm{OR}_{TY}(C=0)}$, respectively (see Table VI.5).

Calculation of prior mean β_0^{mean} and prior standard deviation β_0^{sd} for β_0, prior mean β_X^{mean} and prior standard deviation β_X^{sd} for β_X, and prior correlation ρ_{0X} between β_0 and β_X

i. For 10,000 iterations

a. Random draw of λ from $N\left(\lambda^{mean}, \lambda^{sd}\right)$

b. Random draw of ε from $N\left(\varepsilon^{mean}, \varepsilon^{sd}\right)$

c. Random draw of α_0 from $N\left(\alpha_0^{mean}, \alpha_0^{sd}\right)$

d. Computation of $Se = 0.5 + 0.5\text{expit}(\lambda)$

e. Computation of $Sp = 0.5 + 0.5\text{expit}(\varepsilon)$

f. Computation of

$$\beta_0 = \log\left(\frac{(1-Se)\times P(T=1|Y=0,C=0)}{Sp\times(1-P(T=1|Y=0,C=0))}\right) = \log\left(\frac{(1-Se)}{Sp}\right) + \alpha_0$$

g. Computation of $\beta_X = \log\left(\frac{Se\times Sp}{(1-Se)\times(1-Sp)}\right)$

ii. Computation of β_0^{mean}, β_0^{sd}, β_X^{mean}, β_X^{sd}, ρ_{0X} from the 10,000 values obtained for β_0 and β_X

Calculation of prior mean β_Y^{mean} and prior standard deviation β_Y^{sd} for β_Y

$$\beta_Y^{mean} = \log\left(OR_{TY}(C=0)^{mean}\right)$$

$$\beta_Y^{sd} = \frac{\beta_Y^{upper} - \beta_Y^{lower}}{2\times 1.96},$$

where $\beta_Y^{lower} = \log\left(OR_{TY}(C=0)^{lower}\right)$ and $\beta_Y^{upper} = \log\left(OR_{TY}(C=0)^{upper}\right)$

Calculation of prior mean β_C^{mean} and prior standard deviation β_C^{sd} for β_C

$$\beta_C^{mean} = \log\left(OR_{TC}\left(Y=0\right)^{mean}\right)$$

$$\beta_C^{sd} = \frac{\beta_C^{upper} - \beta_C^{lower}}{2\times 1.96},$$

where $\beta_C^{lower} = \log\left(OR_{TC}\left(Y=0\right)^{lower}\right)$ and $\beta_C^{upper} = \log\left(OR_{TC}\left(Y=0\right)^{upper}\right)$

Calculation of prior mean β_{YC}^{mean} and prior standard deviation β_{YC}^{sd} for β_{YC}

$$\beta_{YC}^{mean} = \log\left(\frac{OR_{TY}\left(C=1\right)^{mean}}{OR_{TY}\left(C=0\right)}\right)$$

$$\beta_{YC}^{sd} = \frac{\beta_{YC}^{upper} - \beta_{YC}^{lower}}{2\times 1.96},$$

where $\beta_{YC}^{lower} = \log\left(\frac{OR_{TY}\left(C=1\right)^{lower}}{OR_{TY}\left(C=0\right)}\right)$ and $\beta_{YC}^{upper} = \log\left(\frac{OR_{TY}\left(C=1\right)^{upper}}{OR_{TY}\left(C=0\right)}\right)$

Appendix III – Program codes

Appendix AIII.1. R code for the implementation of Semi-Bayes adjustment towards the global mean (see Chapters II, III, IV, V)

#This function runs Semi-Bayes adjustment towards the global mean for groups of estimates. An "input" folder must be created under the working directory (here C:/Documents *) and for each group of estimates within which Semi-Bayes adjustment needs to be perform, a .txt file must be created under the "input" folder containing 3 columns: The first one with variable names (e.g. job or industries codes), the second one with ML beta estimates and the third one with the variances of the ML estimates.*

```
SemiBayes <- function(directory,vart){

        setwd(directory)

        #Returns the names of all the files of the "input" directory.
        file<-c(list.files("input"))
        file<-as.character(file)

        #Creates the names of the output files
        outputsb<-paste("SB_",file,sep="")

        for (i in 1:length(file)){
                Data<-read.table(paste("input/",file[i],sep=""),header=TRUE,sep="\t",
dec=".",colClasses=c("character","numeric","numeric"),strip.white=FALSE)
                Data <- na.omit(Data)
                bhat<-as.numeric(c(Data[[2]]))
                vhat<-as.numeric(c(Data[[3]]))
                name<-as.character(Data[[1]])
                newbeta<-vector(length=length(bhat))
                newvar<-vector(length=length(bhat))
                SB_OR<-vector(length=length(bhat))
                SB_OR_LL<-vector(length=length(bhat))
                SB_OR_UL<-vector(length=length(bhat))
                n <- length(bhat)
                w <- 1/(vhat + vart)
                pi <- sum(w*bhat)/sum(w)
                D <- bhat - pi
                varo <- sum(w*D*D)/sum(w)
                varm <- sum(w*vhat)/sum(w)
                newbeta<- w*((vart*bhat)+(vhat*pi))
                E <- w* vhat*D*sqrt(varm/vhat)
                newvar <- vhat * (1-vhat*w) + (2*E*t(E))/n
                SB_OR<-c(exp(newbeta))
                SB_OR_LL<-c(exp(newbeta-1.96*sqrt(newvar)))
                SB_OR_UL<-c(exp(newbeta+1.96*sqrt(newvar)))
                newstd<-t(sqrt(newvar))
```

```
                out<-cbind(name,newbeta,newstd,SB_OR,SB_OR_LL,SB_OR_UL)
                b<-as.data.frame(out)
                names(b)[c(1)] <- c("Var")
                names(b)[c(2)] <- c("SBbeta")
                names(b)[c(3)] <- c("SBstd")
                names(b)[c(4)] <- c("SBOR")
                names(b)[c(5)] <- c("SBlower")
                names(b)[c(6)] <- c("SBupper")

                write.table(b, paste("",outputsb[i],sep=""), sep="\t", col.names=TRUE,
row.names=FALSE, quote=F, na="NA")
        }
}


#Runs the function SemiBayes
SemiBayes("C:/Documents",0.25)
```

Appendix AIII.2. R code for the implementation of Bayesian hierarchical regression (see Chapter V)

```
#This function runs Bayesian hierarchical regression.
#betas is a 3-column matrix (1st column=variable names (e.g. ISCO codes), 2nd column=1st stage-coefficients ML estimates, 3rd column=1st stage-coefficients ML estimates standard errors)
# Z is the Z matrix
#Z_weights is a 3-column matrix used to compute the second-stage covariance matrix(1st column=Z.11+2*Z.12, 2nd column=Z.21+2*Z.22, 3rd column=Z.31+2*Z.32)
# tau is the shrinkage parameter tau

hr<-function(directory,betas,Z,Z_weights,tau){
  setwd(directory)
  ilo<-as.character(betas[,1])
  beta<-betas[,2]
  se<-betas[,3]
  Z<-as.matrix(Z)
  Z_weights<-as.matrix(Z_weights)
  # replace NAs with 0
  Z<-replace(Z,is.na(Z),0)
  Z_weights<-replace(Z_weights,is.na(Z_weights),0)
  rowLength<-length(ilo)

  # Computes the diagonal terms of T
  shrinkage<-1/exp(rowMeans(Z_weights))
  # Computes the diagonal terms of the second-stage covariance matrix
  var_resid<-tau^2*shrinkage
  hist(var_resid)

  # Computes the W matrix
  wt<-(se[]^2+tau^2*shrinkage)^-1

# Fits the second-stage model, estimates the second-stage coefficients and saves them in a .txt file
  lmsummary<-summary(lm(beta~Z,,,weights=wt))
  outputVector<-c(lmsummary$coefficients[,1],lmsummary$coefficients[,2])
write.table(matrix(outputVector,length(Z[1,])),file=paste("hr_tau",tau,"_pi.txt",sep=""),
col.names=FALSE,row.names=FALSE,quote=FALSE)

  # Computes the prior means for the first-stage coefficients
  second.beta<-Z %*% lmsummary$coefficients[,1]
  Z_transpose<-t(Z)
```

```
  # Computes the hierarchical regression estimates
  B<-(se^2)*wt
  hr.beta<-B*second.beta + (1-B)*beta

  # Computes the covariance matrix of the prior means for the first-stage coefficients
and the covariance matrix of the hierarchical regression estimates
  zprimewmatrix<-apply(as.matrix(Z_transpose),1,function(x,weights)
x*weights,weights=wt)
  secondVarFirstHalf<-Z %*% solve(t(zprimewmatrix) %*% Z)
  second.Var<-apply(t(secondVarFirstHalf)*Z_transpose,2,sum)
  second.SE<-sqrt(second.Var)
  H<-second.Var * wt
  C<-se^2 * (1- (1-H) * B )
  hr.SE<-sqrt(C)

  # Saves the prior means, the hierarchical regression estimates and their respective
standard errors in a .txt file
  outputVector<-
c("ilo",ilo,"hr_beta",hr.beta,"hr_se",hr.SE,"prior_beta",second.beta,"prior_se",second.S
E)
   write.table(matrix(outputVector,rowLength+1),file=paste("hr_tau",tau,"_hrresults.txt",
sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE)
}

#Runs the function hr
betas<-read.table("Q:/betas.txt",header=TRUE,sep="\t",dec=".",strip.white=TRUE)

Z<-read.table("Q:/Z.txt",header=TRUE,sep="\t",dec=".",strip.white=TRUE)

Z_weights<-
read.table("Q:/Z_weights.txt",header=TRUE,sep="\t",dec=".",strip.white=TRUE)

tau<-0.23

hr("Q:/",betas,Z,Z_weights,tau)
```

Appendix AIII.3. SAS code for the implementation of Sensitivity-Specificity imputation (SS) (see Chapter VI)

1. Fixed Bias parameter Analysis (FBA)

```
/*This SAS macro runs SS FBA
Misclassdata is the dataset with the misclassified smoking status.
Se and sp are initial fixed values for the sensitivity and the specificity
Results is the output table containing the corrected odds ratio and its 95% confidence interval
Correctedsesp is the table showing which initial values for the sensitivity and the specificity had
to be adjusted because they were not compatible with our data*/

%macro ssfbacov(misclassdata,se,sp,results,correctedsesp);
        /*The macro %analyze is then used by the macro %jack which runs the jackknife
        procedure (step vi)*/
        %macro analyze(data=,out=,num=);
                /*step i*/
                proc logistic data=&data descending COVOUT
                OUTEST=EST(rename=(_type_=stat1 _name_=name1));
                        model smokmis=status01 sex01 status01*sex01/maxiter=5000;
                        %bystmt;
                run;

                data est1;
                        set est;
                        if stat1='PARMS';
                        drop _LINK_ stat1 _STATUS_ NAME1 _LNLIKE_;
                run;

                data est2;
                        set est1;
                        do k=1 to &num;
                                output;
                        end;
                        rename status01=bstatus01 sex01=bsex01
                        status01sex01=bstatus01sex01;
                run;

                proc freq data=&data;
                        table smokmis*status01*sex01/out=frequencies;
                        %bystmt;
                run;
```

```
data mis_pvw;
        merge frequencies est2;
        pistar=exp(intercept+bstatus01*status01+bsex01*sex01+bstatus01sex0
        1*status01*sex01)/(1+exp(intercept+bstatus01*status01+bsex01*sex01
        +bstatus01sex01*status01*sex01));
        maxse=pistar+0.01;
        maxsp=1-pistar+0.01;
run;

proc means data=mis_pvw;
        var maxse maxsp;
        output out=maxsesp(drop=_FREQ_ _TYPE_) max=;
run;

/*step ii*/
data maxsesp1;
        length correctionse correctionsp 3;
        set maxsesp;
        se0=&se;
        sp0=&sp;
        se=max(se0,maxse);
        sp=max(sp0,maxsp);
        if se=se0 then correctionse=0;
        else if se ne se0 then correctionse=1;
        if sp=sp0 then correctionsp=0;
        else if sp ne sp0 then correctionsp=1;
        call symput('se1',se);
        call symput('sp1',sp);
run;

/*step iii */
data mis_pvw;
        set mis_pvw;
        se=&se1;
        sp=&sp1;
        ppv=(se*(pistar+sp-1))/(pistar*(se+sp-1));
        npv=(sp*(se-pistar))/((1-pistar)*(se+sp-1));
run;

/*step iv*/
data freq1;
        set mis_pvw;
        do T=0 to 1;
                if smokmis=0 then do;
                        if T=0 then do;
                                freq=npv*count;
                        end;
                        else if T=1 then do;
                                freq=(1-npv)*count;
                        end;
                end;
                if smokmis=1 then do;
                        if T=0 then do;
                                freq=(1-ppv)*count;
                        end;
```

```
                                        else if T=1 then do;
                                                freq=ppv*count;
                                        end;
                                end;
                                output;
                        end;
                run;

                /*step v*/
                proc logistic data=freq1 outest=or_ss(rename=(_type_=stat2 _name_=name2))
                covout descending;
                        model status01=T sex01 /maxiter=5000;
                        weight freq;
                        %bystmt;
                run;

                data &out;
                        set or_ss;
                        where (stat2='PARMS' and _STATUS_ eq '0 Converged');
                run;
        %mend;

        /*step vi*/
        %inc "H:\misclassification\analyses2012\jack.sas";
        %jack(data=&misclassdata,biascorr=0);

        data &results;
                informat method $4.;
                set jackstat;
                where name eq 'T';
                OR_corr=exp(value);
                lower_corr=exp(alcl);
                upper_corr=exp(aucl);
                se=&se;
                sp=&sp;
                method='SSFBA';
                keep method se sp OR_corr lower_corr upper_corr;
        run;

        data &correctedsesp;
                set maxsesp1;
                keep se0 sp0 se sp correctionse correctionsp;
        run;

%mend;
```

2. Probabilistic Bias Analysis (PBA)

```
/*This SAS macro runs SS PBA
Misclassdata is the dataset with the misclassified smoking status.
Mlambda and slambda are the prior mean and standard deviation for lambda, respectively.
Mepsilon and sepsilon are the prior mean and standard deviation for epsilon, respectively.
Results is the output table containing the corrected odds ratio and its 95% simulation interval.
Sespcheck is the table showing which initial draws for the sensitivity and the specificity had to
be adjusted because they were not compatible with our data*/

%macro sspbacovjkfreq(misclassdata,mlambda,slambda,mepsilon,sepsilon,results,sespcheck);

        /*Temporary table storing the estimates*/
        data bigdata.res;
                length j 3;
                j=.;
                value=.;
                jackmean=.;
                stderr=.;
        run;

        /*Temporary table storing the initil draws for the sensitivity and the specificity which
had to be adjusted*/
        data list_sespcheck;
                length j correctionse correctionsp 3;
                se0=.;
                sp0=.;
                se=.;
                sp=.;
                correctionse=.;
                correctionsp=.;
                j=.;
        run;

        /*The macro %analyze is then used by the macro %jack which runs the jackknife
        procedure (step i.g)*/
        %macro analyze(data=,out=,num=);

                /*step i.a*/
                proc logistic data=&data descending OUTEST=bigdata.EST(drop=_type_
                _name_ _LINK_ _STATUS_ _LNLIKE_ ) noprint;
                        model smokmis=status01 sex01 status01*sex01/maxiter=5000;
                        by %bystmt;
                run;
```

```
data est2;
        length j 3;
        set bigdata.est;
        rename status01=bstatus01 sex01=bsex01
        status01sex01=bstatus01sex01;
        do j=1 to 400;
                do l=1 to &num;
                        output;
                end;
        end;
run;

proc freq data=&data;
        table smokmis*status01*sex01/out=frequencies;
        by %bystmt;
run;

data frequencies1;
        do j=1 to 400;
                do k=1 to numrecsbis;
                        set frequencies point=k nobs=numrecsbis;
                        output;
                end;
        end;
        stop;
        length j 3;
        drop k;
run;

proc sort data=est2;
        by j %bystmt;
run;

/*step i.b*/
data est2;
        retain se0 sp0;
        set est2;
        by j;
        if first.j then do;
                lambda=rand('normal',&mlambda,&slambda);
                epsilon=rand('normal',&mepsilon,&sepsilon);
                se0=0.5+0.5*exp(lambda)/(1+exp(lambda));
                sp0=0.5+0.5*exp(epsilon)/(1+exp(epsilon));
        end;
        /*drop lambda epsilon;*/
run;
```

```
data bigdata.mis_pvw;
        merge frequencies1 est2;
        /* step i.a*/
        pistar=exp(intercept+bstatus01*status01+bsex01*sex01+bstatus01sex0
        1*status01*sex01)/(1+exp(intercept+bstatus01*status01+bsex01*sex01
        +bstatus01sex01*status01*sex01));
        /*step i.c*/
        maxse=pistar+0.01;
        maxsp=1-pistar+0.01;
        drop intercept bstatus01 bsex01 bstatus01sex01;
run;

/*step i.c*/
proc means data=bigdata.mis_pvw nway ;
        where j=1;
        var maxse maxsp;
        output out=maxsesp(drop=_FREQ_ _TYPE_) max=;
run;

data _null_;
        set maxsesp;
        call symput('maxse',maxse);
        call symput('maxsp',maxsp);
run;

data bigdata.mis_pvw1;
        length T 3;
        set bigdata.mis_pvw;
        se=max(se0,&maxse);
        sp=max(sp0,&maxsp);
        /*step i.d*/
        ppv=(se*(pistar+sp-1))/(pistar*(se+sp-1));
        npv=(sp*(se-pistar))/((1-pistar)*(se+sp-1));
        /*step i.e*/
        do T=0 to 1;
                if smokmis=0 then do;
                        if T=0 then do;
                                freq=npv*count;
                        end;
                        else if T=1 then do;
                                freq=(1-npv)*count;
                        end;
                end;
                if smokmis=1 then do;
                        if T=0 then do;
                                freq=(1-ppv)*count;
                        end;
                        else if T=1 then do;
                                freq=ppv*count;
                        end;
                end;
                output;
        end;
run;
```

```
	data sespcheck;
		length correctionse correctionsp 3;
		set est2;
		by j;
		if first.j;
		se=max(se0,&maxse);
		sp=max(sp0,&maxsp);
		if se=se0 then correctionse=0;
		else if se ne se0 then correctionse=1;
		if sp=sp0 then correctionsp=0;
		else if sp ne sp0 then correctionsp=1;
		keep correctionse correctionsp se0 se sp0 sp j;
	run;

	/*step i.f*/
	proc logistic data=bigdata.mis_pvw1 outest=or_sspba(drop=_link_ _type_
	_name_ _lnlike_ sex01 _STATUS_ intercept) descending noprint;
		model status01=T sex01 /maxiter=5000;
		weight freq;
		by j %bystmt;
	run;

	data &out;
		set or_sspba;
	run;
%mend;

/*step i.g*/
%inc "H:\misclassification\analyses2012\nadine\jackknifeprior.sas";

%do m=1 %to 25;
	%put sspba cycle &m;
	%let Time=%sysfunc(time(),time8.0);
	%put &Time;
	%jack(data=&misclassdata,stat=T,id=j,biascorr=0);
	proc append base=bigdata.res data=bigdata.jackstat force;
	run;
	proc append base=list_sespcheck data=sespcheck force;
	run;
%end;

/*step i.h*/
data list_sspba;
	set bigdata.res;
	where value ne .;
	logOR=rand('normal',value,stderr);
run;

/*step ii*/
proc univariate data=list_sspba CIPCTLDF noprint;
	var value;
	output out=univpcts_ss pctlpre=p pctlpts=2.5,50,97.5 n=n mean=meanlogOR;
run;
```

```
	proc univariate data=list_sspba CIPCTLDF noprint;
		var logOR;
		output out=univpcts_ss_rdm pctlpre=p pctlpts=2.5,50,97.5 n=n
		 mean=meanlogOR;
	run;

	data &results;
		informat method $5.;
		set univpcts_ss (in=i1) univpcts_ss_rdm;
		OR_corr=exp(p50);
		lower_corr=exp(p2_5);
		upper_corr=exp(p97_5);
		mlambda=&mlambda;
		slambda=&slambda;
		mepsilon=&mepsilon;
		sepsilon=&sepsilon;
		method='SSPBA';
		if i1 then adjustrandom=0;
		else adjustrandom=1;
		keep method mlambda slambda mepsilon sepsilon OR_corr lower_corr
		 upper_corr meanlogOR adjustrandom;
	run;

	data &sespcheck;
		set list_sespcheck;
		if j ne .;
	run;
%mend;
```

Appendix AIII.4. SAS code for the implementation of Direct Imputation (DI) (see Chapter VI)

1. Fixed Bias parameter analysis (FBA)

```
/*This SAS macro runs DI FBA.
Misclassdata is the dataset with the misclassified smoking status.
Beta0, betaX, betaY, betaC and betaYC are fixed values given to the imputation model
coefficients
Results is the output table containing the corrected odds ratio and its 95% confidence
interval.*/


%macro difbacov(misclassdata,beta0,betaX,betaY,betaC,betaYC,results);

        proc freq data=&misclassdata;
                table smokmis*status01*sex01/out=frequencies;
        run;

        /*step i*/
        data frequencies;
                set frequencies;
                beta0=&beta0;
                betaX=&betaX;
                betaY=&betaY;
                betaC=&betaC;
                betaYC=&betaYC;
                pi1xyc=exp(beta0+betaX*smokmis+betaY*status01+betaC*sex01+betaYC
                *status01*sex01)/(1+exp(beta0+betaX*smokmis+betaY*status01+betaC*sex01
                +betaYC*status01*sex01));
        run;

        /*step ii*/
        data freq1;
                set frequencies;
                do T=0 to 1;
                        if T=0 then do;
                                freq=(1-pi1xyc)*count;
                        end;
                        else if T=1 then do;
                                freq=pi1xyc*count;
                        end;
                        output;
                end;
        run;
```

```
        /*step iii*/
        proc logistic data=freq1 outest=or_di(rename=(_type_=stat2 _name_=name2)) covout
        descending;
                model status01=T sex01 /maxiter=5000;
                weight freq;
        run;

        PROC TRANSPOSE DATA=or_di OUT=tordi;
                VAR T ;
                id name2;
        run;
        data &results;
                set tordi;
                estimate=status01;
                variance=T;
                OR_corr=EXP(estimate);
                lower_corr= EXP(estimate - 1.96*SQRT(variance));
                upper_corr= EXP(estimate + 1.96*SQRT(variance));
                method='DIFBA';
                beta0=&beta0;
                betaX=&betaX;
                betaY=&betaY;
                betaC=&betaC;
                betaYC=&betaYC;
                keep method beta0 betaX betaY betaC betaYC OR_corr lower_corr upper_corr;
        run;

%mend;
```

2. Probabilistic Bias Analysis (PBA)

```
/*This SAS macro runs DI PBA.
Misclassdata is the dataset with the misclassified smoking status.
Mean0,meanX,meanY,meanC and meanYC are the prior means for beta0, betaX, betaY, betaC
and betaYC, respectively
Sd0,sdX,sdY,sdC and sdYC are the prior standard deviations for beta0, betaX, betaY, betaC and
betaYC, respectively.
Corr0X is the prior correlation between beta0 and betaX.
Stderrmis is the standard error of the logOR obtained with the misclassified smoking status.
Results is the output table containing the corrected odds ratio and its 95% simulation interval.*/

%macro dipbacovcorr(misclassdata,mean0,meanX,meanY,meanC,meanYC,sd0,sdX,sdY,sdC,
sdYC,corr0X,stderrmis,results);

        /*Temporary table storing the estimates*/
        data list_di;
                estimate=.;
                boot_parm=.;
        run;

        %do m=1 %to 100;
                %put &m;
                %let Time=%sysfunc(time(),time8.0);
                %put &Time;

                data outboot3/view=outboot3;
                        length j 3;
                        do j=1 to 1000;
                                do _i=1 to numrecs;
                                        set &misclassdata point=_i nobs=numrecs;
                                        output;
                                end;
                        end;
                        stop;
                run;

                data outboot4;
                        set outboot3;
                        by j;
                        retain beta0 betaX betaY betaC betaXY betaXC betaYC;
                        /*step i.a*/
                        if first.j then do;
                                mean0=&mean0;
                                meanX=&meanX;
                                meanY=&meanY;
                                meanC=&meanC;
                                meanYC=&meanYC;
                                sd0=&sd0;
                                sdX=&sdX;
                                sdY=&sdY;
                                sdC=&sdC;
                                sdYC=&sdYC;
                                corr0X=&corr0X;
                                beta0=rand('normal',0,1);
```

```
                betaX=corr0X*beta0+sqrt(1-corr0X**2)*rand('normal',0,1);
                beta0=mean0+sd0*beta0;
                betaX=meanX+sdX*betaX;
                betaY=rand('normal',meanY,sdY);
                betaC=rand('normal',meanC,sdC);
                betaYC=rand('normal',meanYC,sdYC);
            end;
            /*step i.b*/
            pi1xy=exp(beta0+betaX*smokmis+betaY*status01+betaC*sex01+
            betaYC*status01*sex01)/(1+exp(beta0+betaX*smokmis+betaY*status0
            1+betaC*sex01+betaYC*status01*sex01));
            /*step i.c*/
            T=rand('bernoulli',pi1xy);
        run;

        /*step i.d*/
        proc logistic data=outboot4 outest=or_di (drop=_link_ _type_ _name_ _lnlike_
        sex01 intercept) descending noprint;
            model status01=T sex01 /maxiter=5000;
            by j;
        run;

        /*step i.e*/
        data or_di1;
            set or_di;
            if _STATUS_ eq '0 Converged';
            estimate=T;
            znor=rand('normal',0,1);
            stderrmis=&stderrmis;
            boot_parm=estimate-znor*stderrmis;
            keep estimate boot_parm;
        run;

        proc append base=list_di data=or_di1 force;
        run;

%end;

data list_di;
        set list_di;
        if boot_parm ne .;
run;

/*step ii*/
proc univariate data=list_di CIPCTLDF noprint;
        var estimate;
        output out=univpcts_di pctlpre=p pctlpts=2.5,50,97.5 n=n mean=meanlogOR;
run;

proc univariate data=list_di CIPCTLDF noprint;
        var boot_parm;
        output out=univpcts_di_rdm pctlpre=p pctlpts=2.5,50,97.5 n=n
        mean=meanlogOR;
run;
```

```
	data &results;
		informat method $5.;
		set univpcts_di(in=i1) univpcts_di_rdm;
		OR_corr=exp(p50);
		lower_corr=exp(p2_5);
		upper_corr=exp(p97_5);
		mean0=&mean0;
		meanX=&meanX;
		meanY=&meanY;
		meanC=&meanC;
		meanYC=&meanYC;
		sd0=&sd0;
		sdX=&sdX;
		sdY=&sdY;
		sdC=&sdC;
		sdYC=&sdYC;
		corr0X=&corr0X;
		method='DIPBA';
		if i1 then adjustrandom=0;
		else adjustrandom=1;
		keep method mean0 meanX meanY meanC meanYC sd0 sdX sdY sdC sdYC
		corr0X OR_corr lower_corr upper_corr meanlogOR adjustrandom;
	run;

%mend;
```

Appendix AIII.5. WinBugs code for the implementation of MCMC analyses 1 and 2 (see Chapter VI)

1. MCMC analysis 1 (set of priors 1)

```
#model#
model {
        a00~dbin(p00,N00)
        a01~dbin(p01,N01)
        a10~dbin(p10,N10)
        a11~dbin(p11,N11)
        p00<-pi00*Se+(1-pi00)*(1-Sp)
        p01<-pi01*Se+(1-pi01)*(1-Sp)
        p10<-pi10*Se+(1-pi10)*(1-Sp)
        p11<-pi11*Se+(1-pi11)*(1-Sp)
        Se<-(1+exp(lambda)/(1+exp(lambda)))/2
        Sp<-(1+exp(epsilon)/(1+exp(epsilon)))/2
        lambda ~ dnorm(m.lambda, prec.lambda)
        epsilon ~ dnorm(m.epsilon, prec.epsilon)
        logit(pi00)<-alpha0
        logit(pi01)<-alpha0+alphaC
        logit(pi10)<-alpha0+alphaY
        logit(pi11)<-alpha0+alphaY+alphaC+alphaYC
        alpha0~dnorm(0,0.25)
        alphaC~dnorm(0,0.0625)
        alphaY~dnorm(0,0.44)
        alphaYC~dnorm(0,0.11)
        ORadj<-(N11*pi11*N01*(1-pi01)/(N11+N01)+N00*(1-
pi00)*N10*pi10/(N10+N00))/(N11*(1-pi11)*N01*pi01/(N11+N01)+N10*(1-
pi10)*N00*pi00/(N10+N00))
}


#data#
list(N00=353, N01=425, N10=229, N11=226, a00=143, a01=220, a10=158, a11=171,
m.lambda=-0.41, m.epsilon=0.41, prec.lambda=4, prec.epsilon=4)

#initial values#
list(lambda=-0.5,epsilon=0, alpha0=-0.5, alphaC=0.5, alphaY=1, alphaYC=0.5)
list(lambda=0, epsilon=0.5, alpha0=0, alphaC=0, alphaY=0, alphaYC=0)
```

2. MCMC analysis 2 (set of priors 1)

```
#model#
model {
        a00~dbin(p00,N00)
        a01~dbin(p01,N01)
        a10~dbin(p10,N10)
        a11~dbin(p11,N11)
        p00<-pi00*Se+(1-pi00)*(1-Sp)
        p01<-pi01*Se+(1-pi01)*(1-Sp)
        p10<-pi10*Se+(1-pi10)*(1-Sp)
        p11<-pi11*Se+(1-pi11)*(1-Sp)
        Se<-(1+exp(lambda)/(1+exp(lambda)))/2
        Sp<-(1+exp(epsilon)/(1+exp(epsilon)))/2
        lambda ~ dnorm(m.lambda, prec.lambda)
        epsilon ~ dnorm(m.epsilon, prec.epsilon)
        logit(pi00)<-alpha0
        logit(pi01)<-alpha0+alphaC
        logit(pi10)<-alpha0+alphaY
        logit(pi11)<-alpha0+alphaY+alphaC+alphaYC
        alpha0~dnorm(-0.39,204.08)
        alphaC~dnorm(0.1,8.16)
        alphaY~dnorm(1.94,2.04)
        alphaYC~dnorm(0.46,8.16)
        ORadj<-(N11*pi11*N01*(1-pi01)/(N11+N01)+N00*(1-
pi00)*N10*pi10/(N10+N00))/(N11*(1-pi11)*N01*pi01/(N11+N01)+N10*(1-
pi10)*N00*pi00/(N10+N00))
}


#data#
list(N00=353, N01=425, N10=229, N11=226, a00=143, a01=220, a10=158, a11=171,
m.lambda=-0.41, m.epsilon=0.41, prec.lambda=4, prec.epsilon=4)

#initial values#
list(lambda=-0.5,epsilon=0, alpha0=-0.5, alphaC=-0.5, alphaY=0.6, alphaYC=0)
list(lambda=0, epsilon=0.5, alpha0=-0.28, alphaC=0.5, alphaY=3, alphaYC=1)
```

Appendix IV – Statements of contribution to doctoral thesis containing publications

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Marine Corbin

Name/Title of Principal Supervisor: Professor Neil Pearce

Name of Published Research Output and full reference:

Corbin M, McLean D, 't Mannetje A, Dryson E, Walls C, McKenzie F, Maule M, Cheng S, Cunningham C, Kromhout H, Blair A, Pearce N. Lung cancer and occupation: a New Zealand Cancer Registry-based case-control study. American Journal of Industrial Medicine 2011; 54: 89-101.

In which Chapter is the Published Work: Chapter III

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 80%

  and / or
- Describe the contribution that the candidate has made to the Published Work:

Marine Corbin — Digitally signed by Marine Corbin DN: cn=Marine Corbin, o=Massey University, ou=CPHR, email=m.corbin@massey.ac.nz, c=NZ Date: 2013.10.15 13:53:52 +13'00'

Candidate's Signature — 15/10/2013 — Date

Neil Pearce — Digitally signed by Neil Pearce DN: cn=Neil Pearce, o, ou, email=n.e.pearce@massey.ac.nz, c=NZ Date: 2013.11.07 12:25:56 Z

Principal Supervisor's signature — 07/11/2013 — Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Marine Corbin

Name/Title of Principal Supervisor: Professor Neil Pearce

Name of Published Research Output and full reference:

Richiardi, L., Corbin, M., Marron, M., Ahrens, W., Pohlabeln, H., Lagiou, P., . . . Merletti, F. (2012). Occupation and risk of upper aerodigestive tract cancer: the ARCAGE study. International Journal of Cancer, 130(10), 2397-2406. doi: 10.1002/ijc.26237

In which Chapter is the Published Work: Chapter IV

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  The candidate carried out the statistical analyses, and contributed to the interpretation of the results and manuscript writing.

Marine Corbin — Digitally signed by Marine Corbin DN: cn=Marine Corbin, o=Massey University, ou=CPHR, email=m.corbin@massey.ac.nz, c=NZ Date: 2013.11.07 15:01:15 +13'00'

Candidate's Signature — 07/11/2013 — Date

Neil Pearce — Digitally signed by Neil Pearce DN: cn=Neil Pearce, o, ou, email=n.e.pearce@massey.ac.nz, c=NZ Date: 2013.11.07 12:25:07 Z

Principal Supervisor's signature — 07/11/2013 — Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Marine Corbin

Name/Title of Principal Supervisor: Professor Neil Pearce

Name of Published Research Output and full reference:

Corbin M, Richiardi L, Vermeulen R, Kromhout H, Merletti F, Peters S, Simonato L, Steenland K, Pearce N, Maule M. Hierarchical regression for multiple comparisons in a case-control study of occupational risks for lung cancer. PLoS One 2012; 7: e38944. doi: 10.1371/journal.pone.0038944.

In which Chapter is the Published Work: Chapter V

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 90%

  and / or
- Describe the contribution that the candidate has made to the Published Work:

Marine Corbin — Digitally signed by Marine Corbin DN: cn=Marine Corbin, o=Massey University, ou=CFHR, email=m.corbin@massey.ac.nz, c=NZ Date: 2013.10.15 14:12:03 +13'00'

Candidate's Signature

15/10/2013

Date

Neil Pearce — Digitally signed by Neil Pearce DN: cn=Neil Pearce, o, ou, email=n.e.pearce@massey.ac.nz, c=NZ Date: 2013.11.07 12:25:29 Z

Principal Supervisor's signature

07/11/2013

Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Marine Corbin

Name/Title of Principal Supervisor: Professor Neil Pearce

Name of Published Research Output and full reference:

Corbin M, Pearce N, Maule M, Greenland S. Application of Sensitivity-Specificity imputation and Direct Imputation methods to adjust misclassification of the smoking status in the association between smoking and lung cancer - This paper was ready for submission on the date of submission of the thesis , but the camdidate was waiting for final approval from all authors for the final minor changes before formally submitting it –

In which Chapter is the Published Work: 6

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 80%

  and / or
- Describe the contribution that the candidate has made to the Published Work:

Marine Corbin — Digitally signed by Marine Corbin DN: cn=Marine Corbin, o=Massey University, ou=CPHR, email=m.corbin@massey.ac.nz, c=NZ Date: 2013.11.12 20:38:28 +13'00'

Candidate's Signature

12/11/2013

Date

Neil Pearce — Digitally signed by Neil Pearce DN: cn=Neil Pearce, o=Massey University, ou=CPHR, email=n.e.pearce@massey.ac.nz, c=NZ Date: 2013.11.14 14:33:44 Z

Principal Supervisor's signature

14/11/2013

Date