

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**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 Mannelje, 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 Wellington. 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

Abstract.....	i
Author’s declaration	iii
Acknowledgements.....	iv
Table of Contents	vii
List of tables.....	ix
List of figures	xi
List of Abbreviations	xii
SECTION 1. INTRODUCTION, BACKGROUND AND METHODS.....	1
Chapter I. General introduction.....	2
Chapter II. Background and methods	7
A. Background	7
1. Occupational and environmental risk factors for cancer.....	7
2. Statistical issues in the estimation of risks associated with occupational and environmental exposures	11
B. Methods.....	13
1. Introduction to Bayesian inference	13
2. Shrinkage methods.....	21
3. Bayesian methods for the analysis of bias	36
SECTION 2. RANDOM ERROR.....	46
Chapter III. Lung cancer and occupation: A New Zealand cancer registry-based case-control study	47
Chapter IV. Occupation and risk of upper aerodigestive tract cancer: the ARCAGE study	76
Chapter V. Hierarchical regression for multiple comparisons in a case-control study of occupational risks for lung cancer	95
SECTION 3. SYSTEMATIC ERROR.....	117
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.....	118

SECTION 4. DISCUSSION AND CONCLUSIONS	155
Chapter VII. General discussion	156
A. Key findings in occupational epidemiology of lung cancer and upper aerodigestive tract cancer.....	157
B. Bayesian methods to account for random error	161
1. Summary of the approach	161
2. Key findings.....	163
3. Limitations	165
C. Bayesian methods to adjust for systematic error	167
1. Summary of the approach	167
2. Key findings.....	168
3. Limitations	172
D. Future research	173
E. Conclusions	174
REFERENCES.....	176
APPENDICES	190
Appendix I – Publications arising from the work presented in the thesis	192
Appendix II – Further details of methodology	193
Appendix III – Program codes.....	201
Appendix IV – Statements of contribution to doctoral thesis containing publications.....	221

List of tables

Table II.1. The 22 agents, for which exposures are mostly occupational, without considering pesticides and drugs, which are established human carcinogens (Group 1).....	10
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) before and after Bonferroni and Semi-Bayes adjustments. Men.....	31
Table II.3. Characteristics of several quantitative bias analysis techniques.....	45
Table III.1. Characteristics of the study participants.....	55
Table III.2. Odds ratios (OR) and 95% CIs for a priori high risk occupations.....	62
Table III.3. Odds Ratios (OR) and 95% CIs for a priori high risk industries.....	65
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).....	66
Table IV.1. Selected characteristics of cases and controls.....	82
Table IV.2. Selected occupations and industrial branches. Men.....	86
Table IV.3. Selected occupations and industrial branches by cancer site. Men.....	87
Table V.1. Selected characteristics of cases and controls.....	103
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.....	104
Table V.3. Descriptive statistics for the distribution of the $\ln(\text{OR})$ 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).....	105
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.....	110

Table VI.1. Odds ratios of lung cancer and respective 95% CIs after the application of MIME.....	122
Table VI.2. Prior distributions on sensitivity and specificity for SS PBA	131
Table VI.3. Fixed values for model (2) coefficients in DI FBA	133
Table VI.4. Definition of model (2) coefficients for DI FBA.....	134
Table VI.5. Prior distributions on model (2) coefficients for DI PBA.....	137
Table VI.6. Definition of model (3) coefficients for the fully Bayesian analysis.....	141
Table VI.7.a. Prior distributions for the fully Bayesian analysis corresponding to the SS PBA analysis (Table VI.2)	142
Table VI.7.b. Prior distributions for the fully Bayesian analysis corresponding to the DI PBA analysis (Table VI.5).....	142
Table VI.8. Prevalences of subjects classified as exposed and non-exposed in strata of Y and C	144
Table VI.9. Smoking-lung cancer odds ratios from SS FBA	146
Table VI.10. Smoking-lung cancer odds ratios from DI FBA.....	147
Table VI.11. Smoking-lung cancer odds ratios from SS PBA.....	148
Table VI.12. Smoking-lung cancer odds ratios from DI PBA.....	149
Table VI.13. Smoking-lung cancer odds ratios from MCMC analysis 1.....	150
Table VI.14. Smoking-lung cancer odds ratios from MCMC analysis 2.....	150
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	169
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.....	171

List of figures

Figure II.1. Likelihood function for the proportion of successes θ , given that we obtain 4 successes in our experiment.....	15
Figure II.2. Illustration of Monte Carlo Integration.....	18
Figure II.3. The rifle example (1) - Illustration of bias and scatter.....	22
Figure II.4. The rifle example (2) - Illustration of shrinkage.....	24
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. Men.....	30
Figure V.1. Kernel density distributions of the $\ln(\text{OR})$ s. Kernel density distributions of the $\ln(\text{OR})$ 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).....	106
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 $\tau = 0.76$ vs. Maximum Likelihood (ML) (A), HR with $\tau = 0.59$ vs. ML (B), HR with $\tau = 0.23$ vs. ML (C) and Semi-Bayes adjustment towards the global mean (SB) vs. ML (D).....	109
Figure VI.1. Description of possible ranges for misclassification parameters	145

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