Hierarchical models in the analysis of trends in prevalence of congenital anomalies and risks associated with first trimester medications.

Alana Cavadino

Barts and the London School of Medicine and Dentistry

Queen Mary University of London

Submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy

September 2017

Statement of originality

I, Alana Cavadino, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:
Date: 19/01/2018

Details of collaboration and publications:

Cavadino A, Prieto-Merino D, Addor MC, Arriola L, Bianchi F, Draper E, . . . Morris JK (2016). Use of hierarchical models to analyze European trends in congenital anomaly prevalence. Birth Defects Res A Clin Mol Teratol, 106, 480-8.

This thesis is dedicated to the memory of my father Neil A. Dey (1946-2015), whose infectious enthusiasm and great sense of fun continue to inspire me every day.

Acknowledgements

I would first and foremost like to express my sincere gratitude to my supervisor Professor Joan Morris for her guidance, patience, wisdom and countless enjoyable discussions we have shared. I could not have asked for a more supportive and encouraging mentor. I would also like to thank my other supervisor, Dr David Prieto, for his valuable advice and insights throughout my PhD. I am grateful to Dr Richard Hooper and Dr Jonathan Myles who examined my PhD progression stages at 9 and 18 months, giving me useful suggestions and encouraging me to continue with this research.

I am grateful to many colleagues and fellow PhD students, both past and present. Thanks to my office-mates at the Wolfson Institute, Maria, Felix and Anna, for the company, stimulating discussions, and for putting up with my mutterings! A very special thank you to Annie, Jane and Vicki (team W5.09) for reading and commenting on sections of my thesis, supporting me throughout with advice from their own PhD experiences, providing many laughs, and above all for their invaluable friendship. I am also grateful to Professor Elina Hypponen, who was my first research mentor at UCL, and helped me find the ambition and confidence to pursue a PhD in the first place.

This work would not have been possible without the financial support of a PhD studentship from the Medical Research Council and the Wolfson Institute of Preventive Medicine. I would also like to thank the EUROCAT registries who allowed their data to be used for this project, and the many people throughout Europe involved in providing and processing information, including affected families, clinicians, health professionals, medical records clerks, and registry staff.

Last, but not least, I want to say a huge thank you to all of my wonderful family and friends. You have provided me with unwavering support, taken an interest in my work, put up with me at times and generally kept me sane and happy throughout the past three years. Finally, a small (but important!) shout out to my delicious friend coffee, without which this thesis would have taken a lot longer to write…

Thank you all.

Funding

This work was supported by the Medical Research Council [Award reference: 1504916] and the Wolfson Institute of Preventive Medicine.

Abstract

Background

Early identification of risk factors, in particular first trimester teratogenic medications, for congenital anomalies (CAs) is essential. Despite similarities between different CAs and between different medications, current surveillance methods in Europe examine each CA and each medication separately. This thesis aims to investigate whether the use of hierarchical statistical methods combining information in groups can improve CA surveillance methods.

Methods

EUROCAT is a European network of population-based CA registries, with EUROmediCAT comprising those registries with additional information on medication use in pregnancy. Trends in CAs from 2003-2012 in 18 EUROCAT registries (n=81,147) were analysed using Poisson regression models considering each CA separately and using hierarchical models combining related subgroups. First trimester medication exposures from 1995-2011 in 13 EUROmediCAT registries (n=15,058) were analysed. Firstly, groupings of medications and/or CAs were considered when determining the statistical significance of each medication-CA combination, using False Discovery Rate (FDR) procedures to adjust for multiple testing. Secondly, Bayesian hierarchical models were applied to directly model the group effects. The Australian classification system for prescribing medicines in pregnancy was used to independently identify “high risk” medications. The number of “high risk” medications identified by the FDR methods and Bayesian models were compared.

Results

For analysis of trends, grouping EUROCAT CA subgroups using hierarchical models did not provide additional information over that obtained from independent analyses of each subgroup. The double FDR method grouping medications by ATC3 level codes performed better than other FDR methods. Use of Bayesian hierarchical models did not produce enough of an improvement to justify the increased effort of implementing such models.

Conclusions

The current EUROCAT methods of analysing each CA separately remain an appropriate method for the detection of potential changes in prevalence of CAs. The double FDR procedure is recommended for use in routine signal detection analyses of CA data.

Table of contents

List of Figures

List of Tables

Commonly used abbreviations

ATC	Anatomical Therapeutic Chemical
AE	Adverse Event
BHM	Bayesian Hierarchical Model
BUGS	Bayesian inference Using Gibbs Sampling
CA	Congenital Anomaly
CHD	Congenital Heart Defect
CI	Confidence Interval
EUROCAT	European Concerted Action on Congenital Anomalies and Twins
EUROmediCAT	European system for evaluation of safety of medication use in pregnancy in relation to the risk of congenital anomalies
FDR	False Discovery Rate
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
ICBDSR	International Clearinghouse for Birth Defects Surveillance and Research
ICD	International Classification of Diseases
JAGS	Just Another Gibbs Sampler
MCMC	Markov Chain Monte Carlo
NBDPN	National Birth Defects Prevention Network
NTD	Neural Tube Defect
PRR	Proportional Reporting Ratio
PCI	Posterior Credible Interval
SR	Spontaneous Reporting

Chapter 1: Thesis background and rationale

1.1. Introduction

This thesis investigates the use of hierarchical statistical methods as an approach to the routine analyses of congenital anomaly (CA) data in two main areas; firstly, the analysis of changes in prevalence of CAs and, secondly, the detection of medications that may potentially increase the risk of specific CAs when used in the first trimester of pregnancy. The current chapter introduces and contextualises the main topics explored in this thesis, including CA surveillance, Bayesian analysis and the use of hierarchical models.

1.2. Surveillance of congenital anomalies

CAs are structural or functional abnormalities that occur before conception or during a foetus's development and are present at birth, whether clinically obvious at that stage or diagnosed later in life. As a leading cause of both foetal and infant death, around 303,000 neonates globally die each year due to CAs, and it is estimated that around 3-6% of births worldwide are affected by a major CA [Parker et al., 2010, World Health Organization, 2016]. For those who survive past the neonatal period, CAs can lead to lifelong chronic illness and disability, and this carries a significant social, public health, and economic burden. These conditions can be a result of harmful environmental factors known as teratogens; they may also be inherited (originating before conception), or can arise from a complex interaction between both genetic and environmental influences. This is a diverse group of disorders, for which known causal factors include maternal age, medication use, family history, and maternal illness [Harris et al., 2017]. However, only approximately half of all CAs can be linked to a specific known cause or risk factor [Christianson et al., 2005, World Health Organization, 2016], and it is therefore essential that continued efforts are made to identify new potential risk factors.

1.2.1. Background to congenital anomaly surveillance

Before the maternal rubella infection was discovered to be teratogenic in the early 1940s [Gregg, 1941], it was widely believed that the foetus was protected from harmful exposures by the placenta. Twenty years later, a widespread epidemic of deformities in children was linked to the drug thalidomide, which was taken as an anti-morning sickness medication during pregnancy by tens of thousands of women [McBride, 1961, Khoury et al., 1994]. The thalidomide tragedy was the first demonstration on such a large scale that medications could be harmful to the foetus. This led to the establishment and strengthening of regulatory bodies, including more structured regulations for the development and control

of medications. Population-based CA registries were also established across the world, in order to facilitate surveillance and research regarding environmental causes. Since CAs are typically rare diseases [EUROCAT Central Registry, 2012], it is necessary to collect information covering births across an extremely large study population of interest in order to have sufficient numbers to perform meaningful statistical analyses. In Europe, there are a number of regional and national registries collecting data on CAs at a population level. Consequently, a European network of population-based registries for the epidemiologic surveillance of CAs (EUROCAT, http://www.eurocat-network.eu/) was formed in 1979, allowing data to be pooled and compared across Europe, and sharing expertise across the different registries and countries involved. EUROCAT surveys over 1.7 million births from 43 registries in 23 countries across Europe, covering around 30% of the European birth population. One of the main aims of EUROCAT is to perform annual monitoring of the birth prevalence of specific CAs and investigation of the occurrence of any increasing trends or clusters of cases [Dolk, 2005]. Other such networks including both population and hospital-based CA registries include the Latin-American collaborative study of congenital malformations (ECLAMC) [Poletta et al., 2014] and the worldwide International Clearinghouse for Birth Defects Surveillance and Research (ICBDSR), which consists of 40 registries worldwide and includes many of the EUROCAT registries [Botto et al., 2006b]. Collaborative networks also exist within countries; for example, the National Birth Defects Prevention Network (NBDPN) publishes studies from both state and regional level data in the US in an annual special issue of the journal *Birth Defects Research Part A [Kirby and Browne, 2016]*. There is some overlap between the different networks, e.g. some registries are members of both EUROCAT and ICBDSR.

1.2.2. Analysis of medication use during pregnancy and the potential for use of congenital anomaly surveillance databases

The first trimester of pregnancy is an essential stage of development for the foetus; during this time, organs undergo critical steps in their development, and most (although not all) non-inherited CAs occur [Sachdeva et al., 2009]. It is known that certain medications can cause CAs when taken during early pregnancy, yet exposure to prescription and over-the-counter medication during pregnancy is common. It can be difficult for pregnant women to avoid medication use for a number of reasons. Estimates from studies in 1995, 2008 and 2012 showed that 40% of pregnancies worldwide and around half of European pregnancies were unplanned [Sedgh et al., 2014]; early pregnancy exposures may therefore occur before a woman is even aware of her pregnancy. Medications are also needed before and

throughout pregnancy for the treatment of chronic diseases such as asthma, depression, diabetes or epilepsy. Furthermore, medication use can be a result of the pregnancy itself, for example in the treatment of severe morning sickness or gestational hypertension. Total avoidance of medication use during pregnancy is therefore often not possible. Furthermore, it may be detrimental to the health of the mother to avoid a medication that is incorrectly labelled as carrying a fetal risk. Therefore, the potential risks to the foetus must be carefully balanced with the health of the mother.

Medication use during pregnancy has, in fact, become increasingly common in recent decades. A study of the use of prescription and over-the-counter medications during pregnancy in the US demonstrated a rise in the average number of 1st trimester medications from 1.5 in 1977 to 2.6 in 2007 [Mitchell et al., 2011]. This study showed that by 2008, approximately 50% of women had reported taking at least one medication during their first trimester. Similar numbers have been reported by studies in other developed countries; for example, in a web-based questionnaire designed to examine prescription medication use in women in Europe, North and South America and Australia, over 80% of women reported taking at least one medication during their pregnancy [Lupattelli et al., 2014]. Despite their widespread use, however, information on the safety of medicines in human pregnancy is often unavailable, particularly for new products. A key reason for this lack of information is that pregnant women are usually excluded from pre-marketing medication safety studies. A review of research protocols submitted to a single institutional review board, for example, found that 90% of submissions involving drug studies specifically excluded pregnant women [Schonfeld et al., 2013]. Another study reviewed phase IV interventional studies posted on one US website and demonstrated that the exclusion of pregnant women from industry-sponsored clinical trials is common, with over 95% of studies assessed excluding pregnant women, despite studying women of "childbearing potential" and not involving a medication classified as possibly teratogenic [Shields and Lyerly, 2013]. At the point of licensing a medication for marketing, therefore, little or nothing is generally known about the safety for its use during pregnancy. One study estimated that over 80% of medications, many of which had been on the market for up to 20 years, had undetermined teratogenic risk for humans [Lo and Friedman, 2002]. Whilst information about reproductive toxicity is sometimes available from animal studies, these can be limited in their ability to predict possible risks of CA in humans [Wilson J. G., 1979]. For example, studies have demonstrated that rodents are not affected by Isotretonic (a treatment for severe acne), but that this substance is highly teratogenic to primates and humans [Nau, 2001]. As such, post-marketing surveillance strategies and studies are

necessary. Approaches to post-marketing surveillance have included teratogen information services [Schaefer et al., 2005, Chambers, 2011], voluntary adverse event reporting systems e.g. the US Food and Drug Administration [US Food and Drug Administration, 2016], registries for exposures to particular medications such as antiretroviral or antiepileptic drugs [Eldridge et al., 1998] and research studies for specific medications or CAs. These post-marketing surveillance strategies have produced important results, but they are generally not ongoing or systematic. To address this lack of systematic and continually updated knowledge, the potential of the EUROCAT network to carry out CA surveillance in Europe regarding medication safety was identified and described in Meijer et al. [2006]. Similarly, a study of routinely collected ICBDSR data suggested that international networks of CA registries could contribute to post-marketing surveillance of the teratogenicity of medications [Lisi et al., 2010]. The authors of these studies concluded that there was substantial opportunity for international CA networks to perform systematic and ongoing post-marketing surveillance of the fetal effects of medications, and that with existing systems and data only limited resources would be required to carry out this additional surveillance work. The EUROmediCAT research project was therefore established, building on the existing EUROCAT network to establish a European system for the evaluation of safety of medication use in pregnancy, in relation to the risk of CAs [de Jong-van den Berg et al., 2011]. EUROmediCAT aims to identify potential "signals" of adverse effects at the earliest possible stage post marketing, and one of its work packages has therefore developed a systematic signal detection method [Luteijn et al., 2016]. This method searches thousands of medication-CA combinations for potential associations, using a False Discovery Rate (FDR) procedure to adjust for multiple testing.

Investigation of potential teratogens is part of the wider field of pharmacovigilance, in which there has been a vast amount of research and many applications of data mining methods to identify adverse drug reactions using spontaneous reporting data [Almenoff June et al., 2005]. Since existing knowledge on the teratogenic effect of medicines used during pregnancy is generally limited, it is difficult to determine a good reference set of known casual associations, on which signal detection methods can be evaluated. Classification systems for prescribing medicines in pregnancy have been used in Australia, Sweden and the US. These include pregnancy labelling regulations and the introduction of risk categories to assign to medications, from those thought to be of "low risk" when taken during pregnancy to those medications known to carry a high risk of permanent damage to the foetus [Sannerstedt et al., 1996]. The Australian system, for example, provides a publicly available database categorising the risks of medicines during pregnancy, which is

developed and sustained by medical and scientific experts [Australian Government Department of Health, 2016]. However, these types of categorisation systems do not identify specific CAs associated with each "high risk" medication, and so there is no "gold standard" for classifying risks according to specific CAs. Therefore, assessment of signal detection methods for CA data remains challenging.

1.3. Motivation and rationale for thesis

Regular and systematic analyses are essential in order to identify changes in the prevalence of any specific CA, or medications that might potentially be harmful to a foetus in its critical developmental period in early pregnancy. EUROCAT and EUROmediCAT provide a wealth of information regarding CAs across many European populations. Their resulting datasets are large, requiring careful handling and analysis. It is important that the statistical models used at this first stage of surveillance are able to make the most of the available data, and to ensure the efficient direction of consequent resources towards the most appropriate areas for further research. Any incorrect conclusions drawn may cause unnecessary stress or confusion in terms of the advice given to pregnant women; as such, statistically significant associations identified need to be confirmed using external data.

Surveillance of CAs is generally performed using defined sets of subgroups, with many studies using the CA subgroups as defined by EUROCAT [EUROCAT, 2011]. These subgroups are coded according to a structured hierarchy, e.g. the nervous system CA group includes Neural Tube Defects (NTDs), which in turn includes the three subgroups spina bifida, anencephaly, and encephalocele. Similarly, the Congenital Heart Defects (CHD) subgroup includes specific CAs such as ventricular septal defect, atrial septal defect, and tetralogy of Fallot. Several subgroups from different body systems are also known to have a common aetiology and are likely to occur together. For example, around 75% of patients with spina bifida also have foot deformities, and a number of features of spina bifida (such as intrauterine positioning and muscle spasticity) are thought to contribute to the development of clubfoot [Broughton et al., 1994, Gunay et al., 2016]. However, despite known relationships and the existing coding hierarchy of CA subgroups, current surveillance methods examine trends in prevalence, clusters, and associations between risk factors and CAs within each subgroup separately [EUROCAT, 2011, Khoshnood et al., 2013, Loane et al., 2013].

In the analysis of medication use during pregnancy, there are known similarities between certain medications, for example in their chemical properties or particular therapeutic uses. Information regarding these factors has informed the classification and coding of

medications into the internationally used Anatomical Therapeutic Chemical (ATC) system, which was developed jointly by the World Health Organisation and the Nordic Council on Medicines [WHO Collaborating Centre for Drug Statistics Methodology, 2011]. The ATC system uses codes with up to seven digits to classify medications into a hierarchical system, such that codes can be aggregated into groups according to shared properties. Table 1.1 shows how the ATC codes classify active substances into groups at five levels. The first level divides medications into 14 main anatomical groups; the next three levels represent further therapeutic/pharmacological/chemical classifications, and the final level gives the chemical substance. For example, Level 1 (ATC1) gives a one-digit code representing the main anatomical group on which the medication works. However, despite the known commonalities among medications, when examining medication risk during pregnancy, each medication and each CA is typically examined separately.

Table 1.1. The classification of active substances into groups at five levels by the Anatomical Chemical Therapeutic (ATC) classification system.

Level	Description	Number of digits in code	Example code	Description of example code
ATC1	Anatomical main group	1	N	Nervous system
ATC2	Therapeutic subgroup	3	N03	Antiepileptics
ATC3	Pharmacological subgroup	4	N03A	Antiepileptics
ATC4	Chemical subgroup	5	N03AG	Fatty acid derivatives
ATC5	Chemical substance	7	N03AG01	Valproic Acid

Since information on known relationships across CAs and medications is not currently being incorporated in CA surveillance analyses, important associations or trends might not be picked up by current methods. However, the grouping of medications or CAs by their potential teratogenic mechanism has been suggested [van Gelder et al., 2010], and the incorporation of information on known relationships may increase the number of true associations that are detected. For example, consider two CAs that are known to have similar aetiologies and which, when considered separately, both show a positive trend that is not statistically significant. It is feasible that when considered together in one analysis (i.e. sharing information between the two CAs) there may be stronger evidence of a trend across both of these CAs. This example highlights how there is potential for CA surveillance methods to combine information from several subgroups simultaneously, such that the analysis of any particular CA might be improved by considering what is happening in related CAs. Natural hierarchies in the ATC system of drug coding may likewise be used to group

similar drugs together in the analysis of medication use during pregnancy. Consider, for example, a medication for which there is a mild (i.e. not strong or convincing) association with a certain CA. It might be useful to also simultaneously consider the relationship of the same CA with medications that have a similar chemical make-up or therapeutic use, and/or what the effect of that particular medication is on other CAs that may be related (e.g. acting on the same part of the body).

Another aspect of routine statistical monitoring for both CA prevalence and medication data is that large numbers of hypotheses are tested simultaneously in both analyses; this issue is known as multiple testing. In any situation where large numbers of hypothesis tests are being performed, there is a high probability that false positive associations will occur when decisions regarding individual hypotheses are based on unadjusted marginal P-values. Any potential medication safety concern or change in prevalence of a CA that is identified by analysis of routine surveillance data will already, of course, be subject to further detailed investigations when evaluating the strength of evidence regarding that particular association. However, what is of interest for this thesis is whether information regarding the relationships between different CAs and/or medications can be incorporated into the analysis at an earlier stage. This information will be used to try and improve the ability to identify true associations whilst limiting the workload involved in following up numerous false positive associations that may arise from testing multiple CAs and medications in each analysis.

Public health importance

It is important to continue to improve and refine methodology used to analyse CA data so that the most appropriate models are used. This will help to ensure that relevant and accurate information is communicated to healthcare providers and expectant mothers, as well as any woman who is considering or trying to become pregnant. Continued and updated identification of medications that may be harmful to the foetus will directly improve patient care. This is particularly relevant to the European population where chronic diseases such as asthma, depression and diabetes are leading causes of mortality and morbidity [Busse, 2010], and there are therefore large numbers of women requiring medication throughout pregnancy to control these conditions.

Summary

In summary, the main motivation for this thesis was to investigate whether knowledge about similarities between medications and/or CAs can provide logical groupings, which

might improve models for routine analysis of (i) trends in CA prevalence and (ii) potential signals of teratogenic medications.

1.4. Use of hierarchical models and Bayesian methods

This section introduces and summarises the key concepts behind two fundamental statistical approaches that are used throughout the thesis: hierarchical models and Bayesian statistics.

1.4.1. What is a hierarchical model?

Information regarding groupings of medications or CAs may be incorporated into analyses using hierarchical models. A hierarchical (or multilevel) model is called such for two reasons. Firstly, the data follows a structure that has some kind of nested hierarchy, for example, CAs clustered within groups according to which body system they belong. Secondly, the model itself has a hierarchy, in that the parameters of one level (e.g. effects of the CA itself) are controlled by those of the next level up in the model (e.g. the effects of the body system that the CA belongs to) [Gelman Andrew and Hill, 2007]. As such, multiple parameters in the model are related by the structure of the problem, and inference about one unobserved quantity can affect inference of another unobserved quantity. Figure 1.1 presents the basic concept of a hierarchical model; on the left of is a non-hierarchical model, where y_i are a set of independent observations treated as draws from a probability distribution. These are used to estimate θ, a set of parameters that define the data generating process for the y_i.

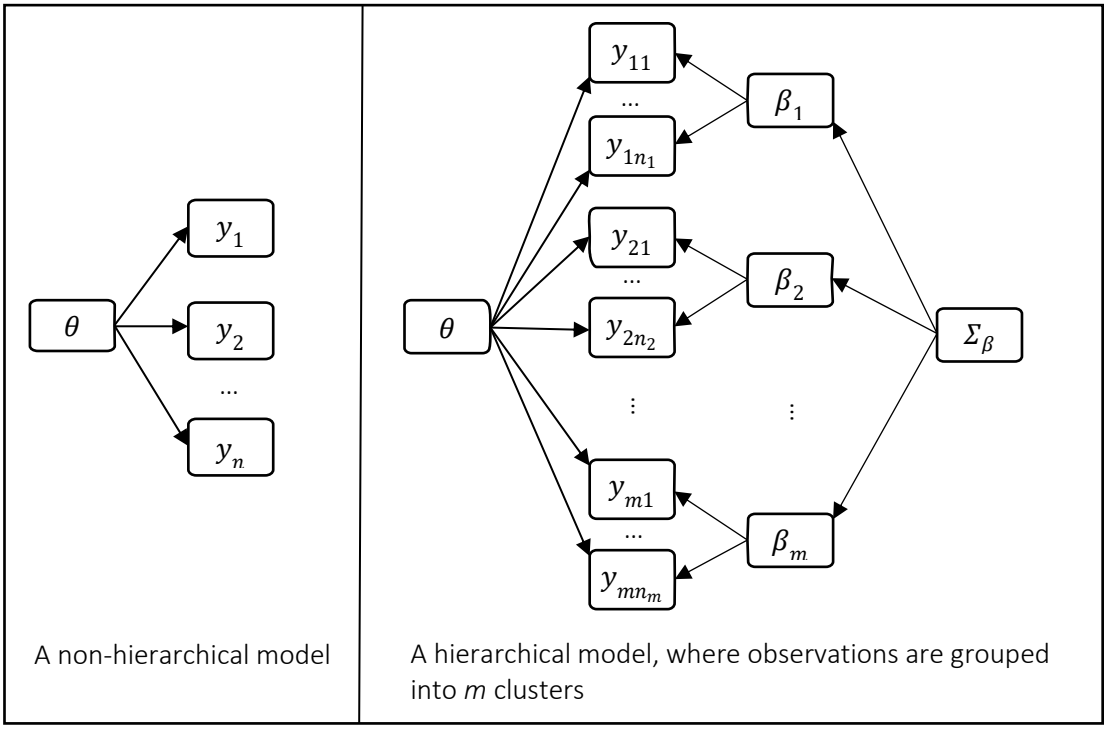


Figure 1.1. Example of a non-hierarchical (left) and a hierarchical (right) model.

On the right hand side of Figure 1.1 is a simple hierarchical model, where the observations y_i now fall into m groups, and the probability distribution over the outcomes is then determined by parameters both across (θ) and within (β_j) these groups. An important difference is that there is now a second probability distribution across the group-specific β parameters, which are generated from a common distribution parameterised by Σ_β. In such a hierarchical model, it is generally expected that observations within any particular group will be more similar to each other compared with observations in other groups. Estimated parameter values are "fixed" if they are shared or constant across the groups (i.e. θ in Figure 1.1), or "random" if allowed to vary across the groups (i.e. the β_j in Figure 1.1).

1.4.2. Frequentist and Bayesian approaches

"Frequentist" statistical inference assumes that observed data are a random repeatable sample, with unknown parameters that are fixed and constant across all potential samples. The Bayesian approach differs in that the observed data are assumed to be fixed, and the model parameters are assumed random. Frequentist inference is therefore based only on the sampling distribution of the observed data, whilst Bayesian inference is based on both the "likelihood" function of the parameters given the observed data, and some "prior" existing information or knowledge about the model parameters. Bayesian statistics has its

foundations in Bayes' theorem, a simple mathematical formula presented by Thomas Bayes in the 18th century that is used to calculate conditional probabilities. Bayes' theorem combines the likelihood and the prior into a "posterior" distribution on which Bayesian inference is based. Bayesian analysis therefore aims to answer questions about unknown parameters using probabilistic statements regarding parameters of the posterior distribution.

Bayesian approaches to reducing the number of errors that arise from multiple testing can lead to lower numbers of false positive findings than frequentist methods, in particular, for data with small cell counts. This is due to "shrinkage", a Bayesian phenomenon wherein results tend to be biased towards the null [Gelman Andrew and Hill, 2007, Kruschke, 2014; section 9.3]. However, this also means that it may be harder to detect some of the true positive associations. In Bayesian Hierarchical Models (BHMs) with multiple levels, the estimates within each level will also demonstrate shrinkage towards the mean estimates for the level; for CA data, this means that the estimates in a group of similar CAs will shrink towards the average estimate across those CAs. In frequentist hierarchical models fixed effects are estimated directly, as they are for standard regression coefficients. Random effects are summarised in terms of their estimated variances, but are not directly estimated. Calculations of the variability around the estimates of random effects for frequentist models require further approaches such as parametric bootstrapping or likelihood profiling, and are therefore not generally reported. The use of simulation-based estimation methods in BHMs offers another advantage over frequentist approaches as estimates are produced for all model parameters, which can be then directly interpreted using their posterior distributions. A measure of uncertainty is thus provided for both individual and group level estimates in the form of a posterior credible interval, which can then be used as a decision rule to determine whether there are credible differences between the individuals and/or the groups [Kruschke and Vanpaemel, 2015]. In general, Bayesian approaches to hierarchical modelling are very flexible, and can avoid many of the challenging approximations and assumptions of frequentist methods for hierarchical modelling [Bolker et al., 2009]. The ability to make direct probability statements from the posterior distribution, combined with the ease of defining models specific to any situation, means that BHMs are a useful approach for many different types of data analysis.

Chapter 2: Thesis aims, objectives and further chapters

2.1. Aim

The overall aim of this thesis was to investigate the practicality and value of hierarchical modelling approaches that group together similar CAs and/or medications in the analysis of CA data.

2.2. Objectives

The specific objectives of this thesis were:

i. To identify and apply suitable BHMs to analyse the change in annual prevalence of several EUROCAT CA subgroups simultaneously.
ii. To compare the performance of models used to fulfil objective (i) to those used currently in annual EUROCAT surveillance programmes, which analyse individual CA subgroups and registries separately.
iii. To identify and apply methods that group together CAs and/or medications when identifying signals of teratogenic medications in EUROmediCAT data, by:
    a) using a post-analysis FDR adjustment for multiple testing that takes groups of medications and/or CAs into account
    b) using BHMs to directly model potential group effects when analysing EUROmediCAT data
iv. To compare the performance of methods used in objective (iii) to each other, and to the signal detection system currently used to analyse EUROmediCAT data.
v. To enable implementation of these methods for use in routine surveillance programs, if they demonstrate improved performance over current methods.

2.3. Outline of thesis chapters

Further chapters in this thesis are summarised as follows

- **Chapter 3** investigates the use of BHMs for the routine analysis of CA prevalence data. An overview of methods currently used are first presented. The EUROCAT prevalence data are then described, and specific groups of CAs that illustrate interesting situations in which grouping CA together may be useful are discussed. BHMs used throughout the chapter are then specified, including details regarding the implementation of Bayesian models using R and JAGS software. Results from these models are then presented and compared to currently used methods of CA surveillance (meeting objectives i and ii).

- **Chapter 4** presents an overview of the use of Bayesian methods in the field of pharmacoepidemiology, and identifies how these might be applied to CA surveillance data. Signal detection methods currently used for CA and medication data are then discussed. Potential improvements to these methods that incorporate information regarding groupings of CAs and/or medications are discussed, and the EUROmediCAT dataset used for these analyses is introduced and described. Finally, the Australian risk categorisation system for prescription of medications in pregnancy is presented, which is used to compare the methods of signal detection investigated in chapters 5 and 6. Potential drawbacks with the use of this categorisation system to compare these methods are highlighted.
- **Chapter 5** presents methods, results and discussion of approaches to EUROmediCAT signal detection that group similar medications and/or CAs together when determining the statistical significance of each test. Various FDR methods that consider groupings in their adjustment for multiple testing are specified and then applied to EUROmediCAT data. Results are compared to existing methods of signal detection for CA data using metrics defined by the Australian risk categorisation system (objectives iiia and iv).
- **Chapter 6** presents methods, results and discussion of approaches to EUROmediCAT signal detection in which information about groups of similar medications and CAs is incorporated using BHMs. Results are compared to those presented in Chapter 5, using metrics defined by the Australian risk categorisation system (objectives iiib and iv).
- **Chapter 7** provides a final discussion and overview of the research presented in this thesis.
- The **appendix** includes:
    - Supplementary material for Chapters 3-6, including: additional information on EUROCAT subgroup coding; examples of R and JAGS scripts; additional details for some methods; full results from BHMs and sensitivity analyses; convergence plots and diagnostics; supplementary figures and tables.
    - A paper published in a peer-reviewed journal based on findings in Chapter 3 [Cavadino et al., 2016]
    - Stata code used to implement the double FDR method presented in Chapter 5

Publications and future work arising from this thesis

A mentioned above, the findings presented in chapter 3 have been published in a peer-reviewed journal [Cavadino et al., 2016]. At the time of submission of this thesis (September 2017), the results from chapters 5 and 6 are being written up as a scientific papers intended for publication in peer reviewed journals. The double FDR method described in chapter 5 is also being implemented in an updated signal detection analysis of EUROmediCAT data, which includes additional registries and recent years of data. These results will also be published separately.

Chapter 3: Analysis of EUROCAT congenital anomaly prevalence data

3.1. Introduction

This chapter begins by describing and assessing the current methods for analysis of trends in CA prevalence data. The use of BHMs that group CAs together in these analyses are discussed and described. These methods are then applied to EUROCAT data to examine whether they offer any improvements over current methods for surveillance of trends in prevalence of CAs.

3.2. Current methods for surveillance of trends in congenital anomaly prevalence

Major CAs are defined by EUROCAT as those that *"require surgical treatment (medical), have serious adverse effects on health or development (functional), or have significant cosmetic impact (cosmetic)"* [EUROCAT Central Registry, 2012]. In Europe, over 80 specific major CAs are monitored systematically and annually by EUROCAT, a network of population-based registries for the epidemiologic surveillance of CAs. A review of the objectives and history of the first 25 years of EUROCAT's CA surveillance is provided by Dolk [2005]. One of the aims of statistical monitoring systems is to enable the identification of unexpected changes or trends in CA prevalence, as these may indicate the presence of a new or unidentified risk factor. This is especially relevant when no specific prior hypotheses have been made regarding potential exposures, such that CAs of potential concern can be identified for which further resources should then be focussed on. CAs flagged in these studies are then subject to more detailed statistical analyses and further investigation by individual registries, to try and determine the potential causes of any observed changes in prevalence. A reported decrease or increase in the prevalence of a CA, for example, may be due to diagnostic or coding changes, and if so this should become apparent upon closer inspection of the data and registries involved. As such, statistical monitoring systems are also used to assess the impact of preventive measures or screening policies. Annual statistical monitoring is hence only one part of CA surveillance; another main aspect is hypothesis-driven studies of more specific risk factors and CAs (where such hypotheses may arise from findings of the hypothesis-generating statistical monitoring analyses).

Analyses of the prevalence of CAs are generally done separately for each CA and each registry; the ICBDSR, for example, produces an annual report that monitors 39 selected CAs separately in each of their member registries, including all member registries submitting

data for that year [ICBDSR, 2014]. In the US, the NBDPN has published national prevalence estimates of selected birth defects using data from 1999–2001 [Canfield et al., 2006], and updated in 2004–2006 [Parker et al., 2010].The NBDPN does not (as yet) perform statistical monitoring across all states in terms of potential changes in CA prevalence. For the analysis of changes in prevalence, EUROCAT produces an annual statistical monitoring report in which both individual registry results and a crude pooling of the data across all registries are presented, with the latest report including data up to the end of 2012 [EUROCAT Central Registry, 2015]. Annual statistical monitoring performed by EUROCAT includes analysis of both five-year "short term" and ten-year "long term" trends, with pooled results providing pan-European estimates of trends in prevalence.

Poisson regression is well suited to count data with rare events and is therefore frequently used to assess trends in CA prevalence [Kirkwood and Sterne, 2003]. Analyses are often done separately for each type of CA, whilst handling of data from the different registries varies, ranging from separate analyses in each registry to a complete pooling of data across all populations. In some studies of specific CAs, for example, registry or country effects have been adjusted for by inclusion as covariates in a regression model [Boyle et al., 2013, Loane et al., 2013] or specified as random effects in multilevel Poisson regression models [Loane et al., 2007]. Random effects Poisson models with linear splines have also been used to model long-term trends in NTDs in EUROCAT registries [Khoshnood et al., 2015]. Methods other than Poisson regression have also been applied in the analysis of CA prevalence data. For example, one study used sequential analysis techniques of cumulative sum and Shewart charts to detect changes in prevalence, and found these useful in identifying changes in reporting and detecting expected increases over time [Babcock et al., 2005]. However, these analyses were also done separately for each type of CA and by four geographical regions included in the study. The Cochran-Armitage test for trends in binomial proportions was also used to identify potential changes in prevalence of 96 CA subgroups over ten years in 21 participating EUROCAT registries [Loane et al., 2011b]. This approach was compared to the use of binomial and Poisson regression in a subset of registries and CAs, which showed agreement in results for 82% of the tests performed, and differences primarily occurring when the observed and/or expected number of cases was below the required minimum for the trend test (defined as at least 5 expected and 2 observed cases per 2-year interval). Although recognised as a "crude" analysis, the authors considered their methods suitable for their purpose of identifying areas requiring further and more careful consideration. This highlights that the priority in statistical monitoring analyses performed by CA consortia is to attain the highest possible detection rates, rather than focussing on

reducing the number of false positive results. The authors also suggest that sophisticated statistical techniques can be difficult to implement as routine screening analyses due to the insufficient numbers of cases that are often present within small registries or for specific types of less common CAs., .

3.3. EUROCAT data

Data for analyses in this chapter come from EUROCAT, a European network of population-based registries established for the epidemiologic surveillance of CAs. In 2012, EUROCAT consisted of 43 multiple-source and high-quality registries in 23 countries (Figure 3.1), including over 1.7 million births per year and covering around 30% of births in the European Union [EUROCAT Central Registry, 2016]. Registries use multiple sources of information to ascertain all CA cases, including live birth, fetal death and termination of pregnancy for fetal anomaly. Data are obtained through a combination of active case finding and voluntary reporting (i.e. case notifications made directly to registries). Sources include hospital discharge lists, maternity, neonatal and paediatric centres, cytogenetic laboratories and fetal ultrasound screening [Greenlees et al., 2011].

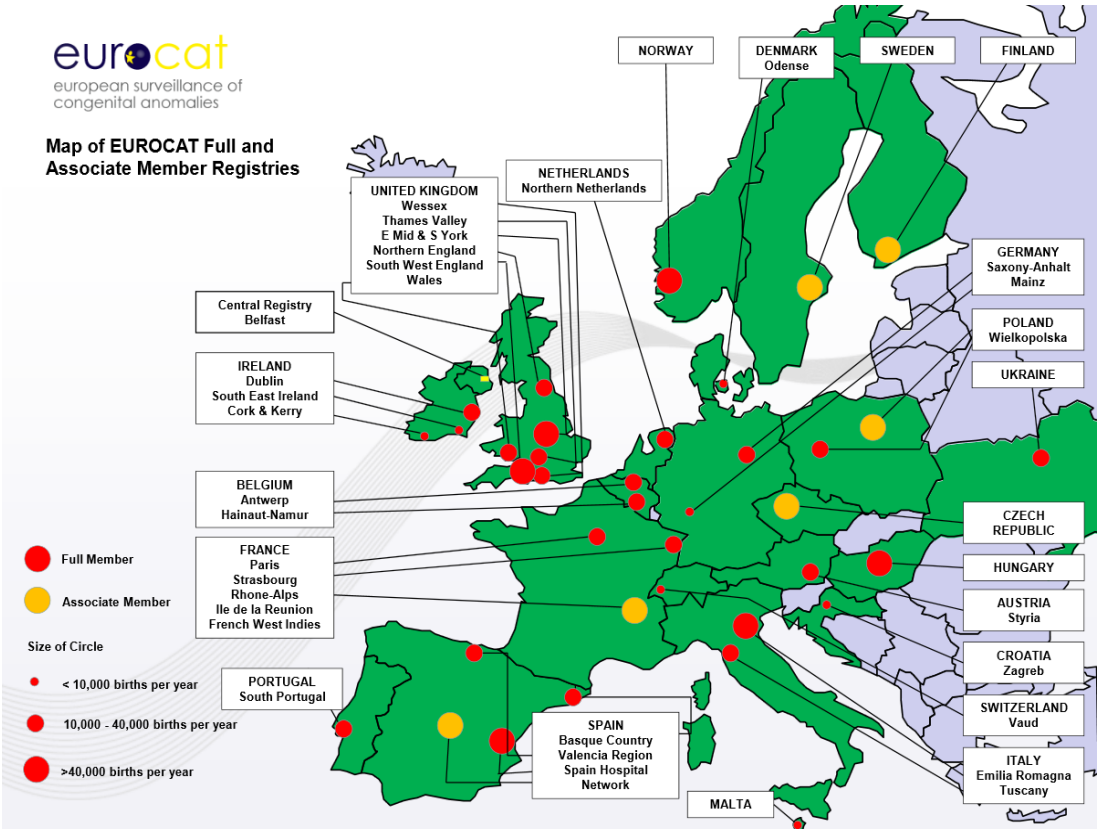


Figure 3.1. Map of EUROCAT Full and Associate Member Registries (as of 2012, http://www.eurocat-network.eu/content/EUROCAT-Map.pdf).

The World Health Organization's International Classification of Diseases (ICD) is a standardised diagnostic classification system that is used worldwide for clinical and research purposes. ICD coding includes a chapter on "Congenital malformations, deformations and chromosomal abnormalities"; however, this often lacks specific or adequate detail for CAs, particularly for genetic syndromes [WHO/CDC/ICBDSR, 2014]. EUROCAT and the ICBDSR have consequently defined modified versions of this coding system such that their data may be analysed for surveillance and research purposes. The EUROCAT data management programme uses the ICD version 9 or 10 (including the British Paediatric Association extension where available, giving a supplementary one-digit extension to ICD-10 codes to allow greater specificity of coding) to assign all major CA cases to EUROCAT CA subgroups, with each case having up to 9 syndrome or malformation codes. EUROCAT produces and maintains standardised malformation coding guidelines and detailed clinical definitions for each EUROCAT subgroup. These enable pooling of data across the different member registries, such that the coding of CAs (which may originally be in the form of written text, for example) is standardised to account for differing levels of accuracy of information in recording. This translation of local to standardised coding requires expert interpretation and knowledge of local conditions for different registries. All coding in the data for this thesis was done according to EUROCAT guide 1.3 [EUROCAT Central Registry, 2005], which uses a hierarchy of codes to classify all cases of non-minor CA into 89 EUROCAT CA subgroups (as of 2012 version of coding; see Appendix Table A1 for details of CA subgroups included in this chapter). EUROCAT anomaly subgroups are grouped in a hierarchical structure, with CAs from the same body system/organ being grouped together. The highest level of EUROCAT coding gives the major organ subgroup, within which there are further classes; for example, spina bifida is in the NTD subgroup, which is within the nervous system group of CAs. A case may be counted only once in each EUROCAT subgroup; however, if a foetus has multiple CAs it can be counted in multiple subgroups. In this thesis, cases with a chromosomal CA as well as any other major (non-chromosomal) CA are only included in the analysis of chromosomal subgroups. Foetuses with a chromosomal or other genetic syndrome are excluded from all other analyses since they are aetiologically different to "non-genetic" CAs.

Data was extracted from the EUROCAT database, including all cases with an expected date of delivery between 1st January 1983 and 31st December 2012. Full and associate member registries with a total prevalence of all CAs of over 2%, available data for at least nine years of the time period from 2003 to 2012 and information available on maternal age according to five-year age groups for the population were included. These restrictions are based on

EUROCAT data quality indicators, which are used to ensure that the data are of sufficient quality [Loane et al., 2011a]. The resulting dataset available for this project included 18 participating registries in 11 countries. The years of data available for each of these registries are presented in Figure 3.2. Many registries started contributing data considerably later than 1983 and only 6 registries had data covering the whole time period. The latest ten years of data were covered by the majority of registries (as per the inclusion criteria), with only 2012 not included for Basque country registry. Ten-year trends have previously been used to reflect recent "long term" trends in European CA prevalence in annual statistical monitoring performed by EUROCAT [Loane et al., 2011b]. Consequently, only data from 2003-2012 were considered for analyses in this chapter.

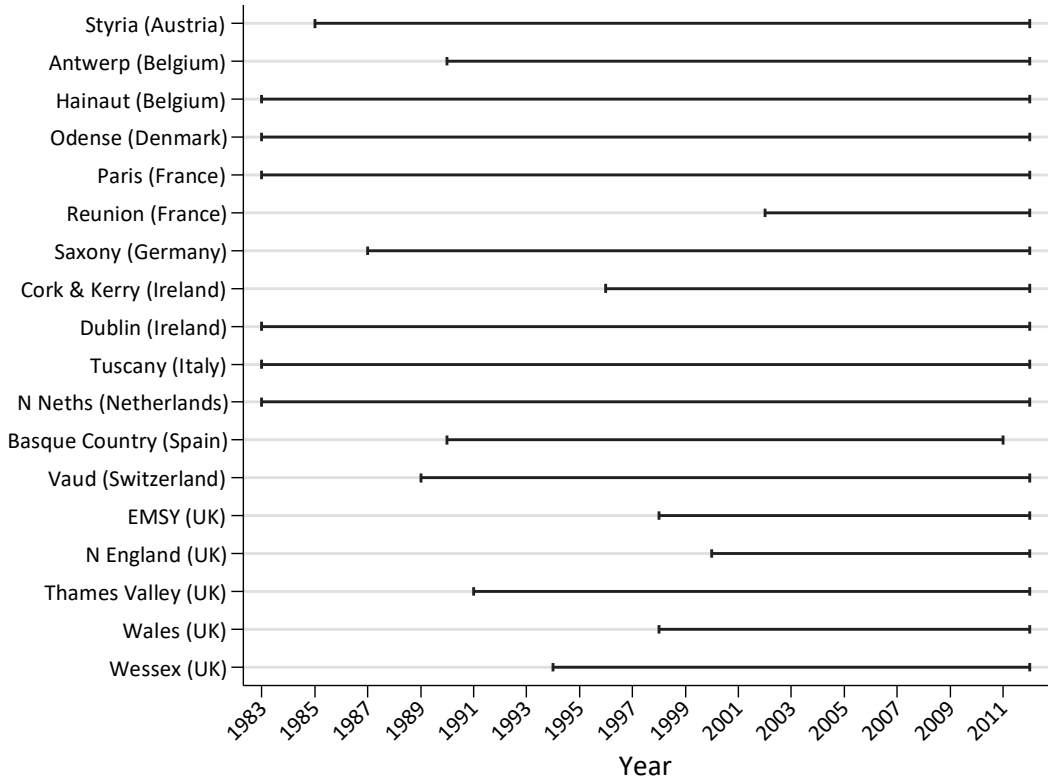


Figure 3.2. Years of available congenital anomaly prevalence data for the 18 EUROCAT registries participating in this study.

3.4. Congenital anomaly subgroups of interest

Hierarchical methods for the analysis of EUROCAT prevalence data were applied to the following groups of CAs; neural tube defects (NTDs), congenital heart defects (CHDs), digestive system and chromosomal anomalies. The reasons these specific subgroups are of interest for these analyses are described below. Further details on EUROCAT coding and brief descriptions of each of the CAs considered in this chapter are in Appendix Table A1.

3.4.1. Neural tube defects

NTDs are a group of malformations that are caused by the incomplete closure of the neural tube during its development, usually within 28 days of gestation [Elwood et al., 1992]. Three defects make up the NTD anomaly subgroups according to EUROCAT coding; spina bifida, encephalocele and anencephaly (subgroup "anencephalus and similar"). Spina bifida is an incomplete closing of the backbone and membranes around the spinal cord, causing an array of lifelong disabilities. Encephalocele is a very rare birth defect in which some part of the skull does not form properly, allowing part of the brain (and membranes that cover it) to protrude through the resulting gap in the skull. Anencephaly is a particularly severe NTD where the upper part of the neural tube does not close and large parts of the brain and skull do not develop at all, causing babies to be stillborn or die shortly after birth [Elwood et al., 1992]. These CAs do sometimes co-occur, in which case EUROCAT coding assigns the most severe NTD subgroup code, such that the spina bifida subgroup excludes any cases with encephalocele or anencephaly, and the encephalocele subgroup excludes any cases with anencephaly [EUROCAT Central Registry, 2005].

The MRC vitamin study established 25 years ago that folic acid supplementation prior to and during early pregnancy can reduce the risk of having a pregnancy with an NTD [MRC Vitamin Study Research Group, 1991]. Over 70 countries worldwide have since introduced mandatory folic acid fortification programmes, including the US, Chile, Canada and Australia. Evidence of significant subsequent declines in NTD prevalence has been demonstrated [Honein et al., 2001, Lopez-Camelo et al., 2005, De Wals et al., 2007, Centers for Disease Control and Prevention, 2010]. However, no European countries have yet introduced mandatory fortification of food with folic acid [Flour Fortification Inititative, 2017]. Despite existing supplementation recommendations and health campaigns across Europe [EUROCAT Central Registry, 2009b], the majority of European women still do not take folic acid supplementation prior to conception [Bestwick et al., 2014]. A report on differing policy and recommendations on folic acid supplementation and fortification across the world concluded that a public health policy of only recommendations alone (i.e. no fortification) does not effectively translate into population-wide declines in NTD rates [Botto et al., 2006a]. There has therefore been considerable interest in NTD prevalence rates across Europe, in particular in comparison to the rest of the world. A recent comprehensive systematic review assessing the current global burden of NTDs found that prevalence estimates vary widely, including within European registries [Zaganjor et al., 2016]. Considerable variation in the size and precision of the estimates of prevalence across registries was therefore expected in an analysis of European trends. However, the NTDs

share similar aetiologies and should be equally sensitive to any changes in folic acid supplementation policies. Long-term trends from 1991 and 2011 of NTDs in Europe have already been assessed in EUROCAT registries, and the total prevalence of NTDs in 2011 was seen to be similar to that in 1991 [Khoshnood et al., 2015]. For this thesis, the effect of modelling the NTD subgroups together was assessed in order to determine whether this provided any additional useful information, regarding recent ten-year trends of a group of CAs where similar overall trends in prevalence are expected.

3.4.2. Chromosomal anomalies

Chromosomal anomalies occur when there is a change in the normal structure or number of chromosomes. Down syndrome, for example, is the presence of an extra full or partial copy of chromosome 21, resulting in a range of mild to moderate developmental disabilities. EUROCAT monitors five chromosomal CAs; Down syndrome (trisomy 21), Edwards syndrome (trisomy 18), Patau syndrome (trisomy 13), Turner syndrome and Klinefelter syndrome. The prevalence of Down, Edwards and Patau syndromes are more or less constant up to a maternal age of around 30; after this age, the prevalence of these anomalies has been demonstrated to increase exponentially before levelling off again at around age 45 [Savva et al., 2010]. The average age of women giving birth in Europe has been steadily rising since around 1980 [Breart, 1997], and increasing maternal ages have thus led to greater numbers of affected pregnancies [Loane et al., 2013]. For these analysis, it was therefore expected that the chromosomal subgroups would show increasing trends if maternal age was not accounted for. This provided a scenario in which a similar trend in each of a group of related CAs was expected, and it was of interest to examine the effect of hierarchical models in this setting. Chromosomal anomalies were also considered as a group including only the three autosomal trisomies (trisomy 21, 18 and 13), for which the relationship with maternal age has been most clearly established [Loane et al., 2013].

3.4.3. Digestive system anomalies

The most recent EUROCAT statistical monitoring report included data up to the end of 2012. In this report, similar increasing trends were found for three of the digestive system anomaly subgroups; oesophageal atresia with or without trachea-oesophageal fistula was a newly identified trend, and duodenal atresia and stenosis and ano-rectal atresia and stenosis were both identified as continuing increasing trends that had been present in the previous years' statistical monitoring report [EUROCAT Central Registry, 2015]. There were a further five digestive system subgroups, for which no significant changes in prevalence were observed. The application of hierarchical models to all of the digestive system CAs

was therefore considered, as an example of a group of CAs where there are known trends in some (but not all) of the subgroups.

3.4.4. Congenital heart defects

CHDs make up the largest group of CAs, accounting for nearly a third of all major CAs [Dolk et al., 2011]. They are a large and heterogeneous group of CAs, which vary widely in terms of their prevalence, severity, morbidity and mortality. EUROCAT coding defines 16 standard CHD subgroups that have previously been grouped using a hierarchical severity ranking according to perinatal mortality rates in non-chromosomal cases. This is formed of three ordered groups from severity I (high perinatal mortality) to severity III (low perinatal mortality) [EUROCAT Central Registry, 2009a]. Severity I indicates CHDs with high perinatal mortality, comprising single ventricle, tricuspid atresia and stenosis, Ebstein's anomaly, hypoplastic left heart and hypoplastic right heart. Severity II includes common arterial truncus, transposition of great vessels, atrioventricular septal defect, tetralogy of Fallot, pulmonary valve atresia, aortic valve atresia/stenosis, coarctation of aorta and total anomalous pulmonary venous return. Severity III indicates low perinatal mortality, including ventricular septal defect, atrial septal defect and pulmonary valve stenosis. Around 5% of all CHD cases are not included in any of the severity categories because they are subtypes of CHD that are not standard EUROCAT subgroups. There is also a EUROCAT anomaly subgroup defined as "severe CHD", which includes all CHD subgroups in severity groups I and II (i.e. the two more severe groupings as described above). In the 2015 EUROCAT statistical monitoring report, increasing trends were found in the severe CHD group as well as in a number of the more specific CHD subgroups, including single ventricle, tetralogy of Fallot and atrioventricular septal defects [EUROCAT Central Registry, 2015]. For the current analyses, it was of interest to see how hierarchical models would perform with a large and heterogeneous group of CAs. A two level hierarchy that additionally included the grouping of CHDs by the severity subgroups was also considered.

3.5. Methods

Prevalence in EUROCAT analyses was calculated as follows, and is presented as prevalence per 10,000 births unless otherwise stated

$$\text{Total Prevalence} = \frac{\text{Number of cases (LB+FD+TOPFA)}}{\text{Number of births (live and still)}} \text{ x } 10{,}000$$

where cases = cases of CA in population, LB = live births, FD = fetal deaths, TOPFA = termination of pregnancy for fetal anomaly and the denominator includes all live and still

births in the population as declared on official birth registrations. Note that terminations of pregnancy for fetal anomaly are included in the numerator but not the denominator, leading to small discrepancies between the two. Such discrepancies, however, are not considered large enough to have an important effect on prevalence. Confidence intervals (CIs) for total prevalence estimates were calculated using the numbers of cases and births with the lower and upper 95% confidence limits defined by the Poisson distribution [Begaud et al., 2005] as follows

$$95\%\ \mathrm{CI} = \left(\frac{\left(\frac{1.96}{2} - \sqrt{cases + 0.02}\right)^2}{births} \text{ x } 10{,}000,\ \frac{\left(\frac{1.96}{2} + \sqrt{cases + 0.96}\right)^2}{births} \text{ x } 10{,}000 \right)$$

Poisson regression models were used to analyse trends in the prevalence of CA data. Methods of analysis currently used for statistical monitoring of trends in prevalence were compared to hierarchical models for groups of CAs detailed in the previous section. The latest ten years of data were assessed in order to reflect the latest "long term" trends in European CA prevalence, as done in annual statistical monitoring performed by EUROCAT. All data management and cleaning was performed using Stata, version 12 [StataCorp, 2011] The different models applied to each group of CAs are specified in detail below.

3.5.1. A Poisson regression model for analysis of prevalence rates

The Poisson distribution is commonly used to model variation in count data, describing the number of occurrences of an event during a specified time period or region [Kirkwood and Sterne, 2003]. For this chapter, prevalence rates were modelled for the number of cases of CAs each year, so the logarithm of the total births (including live and stillbirths) was included as an offset in models to account for the differing population size in each registry and each year, therefore modelling the relative prevalence of disease. If r_t is the prevalence rate per year and p_t the total births per year, then $r_t = \frac{E(count_t)}{p_t} = \frac{\lambda_t}{p_t}$. The classic Poisson regression model for these data was then specified as

$$y_t \sim Poisson(\lambda_t) = Poisson(r_t p_t)$$

With the log link function such that the transformed mean followed a linear model

$$\log(\lambda_t) = \mu_t = \log(p_t) + \beta_0 + \beta_1 x_t$$

The units t specify the year $t = 1, \ldots, T$ for a total period of T years, y_t is the number of CA cases for year t in a process with prevalence rate λ_t and total births (the relative "exposure") p_t. Then $\log(p_t)$ is the model offset, λ_t is the average yearly prevalence and the coefficient $\beta_1$of x_t is the expected log difference in y_t for each additional unit of x_t, i.e.

the estimated average annual change in number of cases (on the log scale). In all models, the predictor variable x_t was centred, such that the intercept β_0 represents the expected log prevalence rate in the mean year across all observations. If this variable were uncentered, the intercept would represent the expected rate in year "0", which is clearly not realistic or interpretable.

Adjustment for multiple testing in Poisson models

A multiple testing correction was applied to adjust the 95% CIs from individual Poisson models to account for multiple testing within each group of related CAs. Adjustments to CIs for individual models were done using a simple Bonferroni correction. In practice this means that instead of a 95% confidence level, a $\left(1-\frac{\alpha}{k}\right)\%$ interval was used, where k is the number of tests in each group and $\alpha = 5\%$ (see Appendix A1 and Table A2 for further details).

Overdispersion in Poisson models

A defining characteristic of the Poisson regression model is that it does not have a scale parameter, but the mean is restricted to be equal to the variance. This restriction often does not hold in practice, with the variance differing from the mean such that any extra between-subject variability beyond that explained by covariates in the model may be unaccounted for. Under- or overdispersed models can be used to account for this additional variability in cases where the conditional variance is substantially different to the conditional mean. A simple comparison of the sample mean and variance of the dependent count variable can give an indication of whether the data are over- or underdispersed. A crude test for dispersion in a generalised linear mixed model (GLMM) can also be performed by comparing the sum of squared Pearson residuals to the residual degrees of freedom, which are assumed to be equal if the data are equi-dispersed [Cameron and Trivedi, 2013]. A simple chi-square test can be performed on the ratio of these values, with the important caveat that due to issues with counting the number of parameters (and therefore degrees of freedom) in a GLMM, such a test should be considered approximate.

Two common approaches to deal with overdispersed data are the use of quasi-Poisson or negative binomial models [Gelman Andrew and Hill, 2007]. In R, the quasi-Poisson model with estimated dispersion parameter can be fitted with the `glm()` function, by setting `family=quasipoisson`. This estimates a scale parameter from the data using a method of moments estimator, but is not maximum likelihood estimation and so model checks such as likelihood ratio tests or assessment of deviance cannot be used. A quasi-Poisson model can

be specified as follows, where estimates for the coefficients are unaffected but the overdispersion parameter ϕ is used to correct the standard errors

$$Y \sim QP(\mu, \phi), \qquad E[Y] = \mu, \qquad Var(Y) = \phi\mu$$

The negative binomial model [Ross and Preece, 1985] allows the rate of events μ to vary across subjects by including a random subject effect in the model. The rate for each subject is then assumed to follow a gamma distribution, with dispersion parameter α

$$Y \sim NB(\mu, \alpha), \qquad E[Y] = \mu, \qquad Var(Y) = \mu + \alpha\mu^2$$

When α=0, the negative binomial distribution coincides with the Poisson distribution. In the quasi-Poisson model, the mean is linearly related to the variance, whilst in the negative binomial model this relationship is quadratic. Gelman Andrew and Hill [2007; section 14.4 & 15.1] describe how a data-level variance component can also be used to directly model overdispersion in a hierarchical model. This has elsewhere been referred to as a generalised Poisson [Martina et al., 2015] or Poisson-lognormal [Elston et al., 2001] model. In this formulation, the Poisson regression model, as described in 3.5.1, is extended by introducing an additional term ε_t to the log link function

$$y_t \sim Poisson(\lambda_t)$$

$$\log(\lambda_t) = \mu_t + \varepsilon_t = \log(p_t) + \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t \sim normal(0, \sigma_\varepsilon^2)$$

The size of the standard deviation σ_ε associated with this new parameter ε_t can give an idea of how overdispersed the model is. When $\sigma_\varepsilon = 0$, the Poisson-lognormal model reduces to a classic Poisson regression model. Functionally, the Poisson-lognormal model is similar to a negative binomial model [Ver Hoef and Boveng, 2007], and the improvement in model fit for the additional term in the model can be assessed using a deviance test.

Models allowing for overdispersion were investigated to determine which type of model was most sensible for use throughout these analyses in this chapter. This was done using the CA subgroup encephalocele, as an example of a rare CA for which accounting for potential overdispersion may be useful. In all other analyses, models continued to be monitored for evidence of departures from dispersion, which could then be investigated in more detail as required.

Zero-inflation in Poisson models

CA prevalence data can potentially contain large numbers of zero counts due to under-recording [Cameron and Trivedi, 2013]. For example, very rare CAs can be difficult to diagnose; if a particular anomaly is seen only occasionally then it may be less likely to be

recognised or correctly diagnosed when it does arise. In addition, there may be an excess of zeros due to the policy of specific registries not to record specific CAs. Issues of under-recording and excess of zero count data have led to the development of zero-inflated [Lambert, 1992] and hurdle [Mullahy, 1986] models. These methods assume that the excess zeros are generated by a separate process from the count values and can therefore be modelled independently. Club foot, for example, might be considered in this context as this CA subgroup has known underreporting due to the policy of specific registries not to record cases of this particular anomaly (for example the Northern England registry), potentially resulting in zero inflated data. However, if a registry does not record a particular CA then this registry does not contribute to the analysis of this CA (since the observed and expected counts are both zero). This is not a clear logical reason for there to be an excess of zeros in the data that could be generated by a separate process from the count values. Zero-inflated models were therefore not considered appropriate for these analyses and were not investigated further.

3.5.2. Bayesian hierarchical models for congenital anomaly data

Hierarchical models

Hierarchical or multilevel models are extensions of regression for data that has a hierarchical or clustered structure, in which model coefficients are allowed to vary according to these clusters or groupings. Clustered data arise when information can be classified into a number of groups (the clusters) based on certain characteristics of the group members, for which individuals within the same group are more similar to each other than to individuals in other groups. For example, children from the same family might be more similar to each other in terms of their physical and/or mental characteristics than children in other families. Hierarchical models allow for such structures in data by including residual components at each level, i.e. at both the child and the family level. The residual variance is split into two parts; the within-family variance (for the child-level residuals) and the between-family variance (for the family-level residuals). The family-level variance represents unobserved family characteristics that may affect the outcomes of the children, leading to correlations between the outcomes of children in the same family. The key difference between classic regression and hierarchical models is the modelling of the variation between the groups of similar individuals. Traditional alternatives for analysing data with a hierarchical structure are complete pooling, which ignores differences between groups, or no pooling, where each group is analysed in a separate model [Gelman Andrew and Hill, 2007]. For this chapter, hierarchical models allowed the grouping of similar CAs

together, essentially re-estimating the trend for each CA by combining information about its own prevalence with the overall information of all other CAs in a group. Each individual estimate incorporates a weighted average of the estimates for all CAs in the group, and estimates are therefore "pulled in" towards the overall group mean. This reduction in the variance of the estimates within a group is an effect known as shrinkage [Gelman Andrew and Hill, 2007, Kruschke, 2014; section 9.3]. Information from other members of the group can remove some of the uncertainty in estimates for group members with limited amount information, for example for rare CAs with small numbers of cases. Statistically significant effects in a non-hierarchical model might also be diminished when the group distribution is taken into account; shrinkage can therefore help in reducing false-positive results.

Hierarchical modelling of count data can be done in a frequentist setting using GLMMs, which combine linear mixed models and generalised linear models in order to handle non-normally distributed data that has a clustered or hierarchical structure. An assumption of GLMMs is that the parameters follow a multivariate normal distribution, such that the sampling distribution of the log likelihood is proportional to a chi-square distribution; tests for P-values of the fixed effects are therefore approximate, and for random effects no such estimates are generally reported. Methods of calculating the variability around the estimates of random effects in GLMMs include parametric bootstrapping and likelihood profiling [Bates D, 2015]. Another approach to inference for GLMMs is the use of Bayesian posterior sampling, which offers several advantages over frequentist methods; see Gelman Andrew and Hill [2007] for a comprehensive discussion about this. Amongst other features, BHMs are highly flexible, they allow direct probability statements to be made about the parameters of interest and can avoid some of the approximations and assumptions (e.g. reliance on the asymptotic approximation) of frequentist methods. For this thesis, therefore, BHMs were used to calculate the variability around the random effects estimates, and bootstrapping and likelihood profiling were not used.

Bayesian models and the use of simulation-based estimation methods

The Bayesian approach to statistical inference considers parameters as random variables that can be characterized by prior distributions based on previous beliefs or prior knowledge about the parameter of interest, combined with the likelihood function of the observed data [Ntzoufras, 2009]. From Bayes' theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Here $p(\theta)$ is the prior probability density summarising prior knowledge about the parameter of interest, $p(y|\theta)$ is the likelihood function of the parameters give the observed data, and $p(\theta|y)$ the posterior density. The marginal likelihood $p(y)$ is the integral of $p(y|\theta)p(\theta)$ over all values of θ, and is regarded as a normalising constant which ensures that $p(\theta|y)$ is a proper density [Congdon, 2007]. Bayes theorem can then be expressed as

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In this equation, the relative influence of the data and the prior beliefs depend on both the "strength" of the data and how informative the prior is, i.e. how much weight is given to the prior belief in terms of its probability distribution. In Bayesian inference, therefore, priors are combined with the observed data to obtain "updated knowledge" in the form of the posterior probability distribution for the parameter of interest as follows

$$p(\theta|data) \propto likelihood(\theta)\ x\ prior(\theta)$$

Calculation of the posterior distribution is done analytically if the combination (multiplication) of the prior distribution and the likelihood form a known distribution, i.e. one that can be summarised and graphed and which inferences can then be drawn from. There are certain well-known combinations of likelihood functions and probability distributions that, when multiplied together, give the same form of posterior distribution as the prior. These are known as conjugate priors [Raiffa and Schlaifer, 1961]. For example, a Normally distributed prior for a mean combined with a Normally distributed likelihood for a mean produces a Normally distributed posterior, where the posterior mean is then a weighted average of the two means, and the posterior precision (the inverse of the variance) is the sum of the two precisions. However, a posterior distribution can also have a more complicated expression if it is not possible to represent prior beliefs using a distribution that is a known conjugate. In such cases, statistics are often difficult to directly calculate and the density function may not be easily drawn. Alternative methods for summarising characteristics of the posterior distribution have therefore been developed, such as asymptotic approximations, numerical integration and sampling methods [Tierney, 1994]. Simulation based sampling or Monte Carlo methods involve the generation of repeated samples that approximate or converge to a "target" posterior distribution. Parameters of interest are then estimated using characteristics of a random sample drawn from the posterior distribution [Congdon, 2007]. Sampling from a distribution can be done using a Markov Chain, which is a sequence of random draws $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$ such that each value depends only on the value of the previous draw

$$f\left(\theta^{(t+1)}\middle|\theta^{(t)},\dots,\theta^{(1)}\right)=f\left(\theta^{(t+1)}\middle|\theta^{(t)}\right)$$

The two most widely used Markov Chain Monte Carlo (MCMC) methods to generate samples from the target distribution are the Metropolis-Hastings algorithm [Hastings, 1970, Chib and Greenberg, 1995] and Gibbs sampling [Geman and Geman, 1984, Casella and George, 1992].

The Metropolis-Hastings algorithm simulates a Markov chain for the parameter of interest θ with target posterior distribution $f(\theta|y)$, and is described by Ntzoufras [2009; page 43] in the following steps

1. Set the initial values $\theta^{(0)}$
2. For t=1,…,T
    (a) Let $\theta=\theta^{(t-1)}$
    (b) Generate new candidate values θ' from a proposal distribution $q(\theta'|\theta)$
    (c) Calculate an acceptance probability

$$\alpha=\min\left(1,\frac{f(\theta'|y)q(\theta|\theta')}{f(\theta|y)q(\theta'|\theta)}\right)$$

    (d) Update $\theta^{(t)}=\theta'$ with probability α; otherwise set $\theta^{(t)}=\theta$

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm that uses the full conditional distribution for the proposal distribution. This always results in an acceptance probability of 1, so the proposed candidate value is always accepted and the sampler moves to a new value at every step. However, since parameters are updated one at a time, this can lead to high autocorrelation, slow convergence and imprecise estimates. This can be particularly problematic when posterior parameters are highly correlated as it can be difficult to change the value of one of the correlated parameters without simultaneously changing the other. A disadvantage of Gibbs sampling, therefore, is that it can sometimes be ineffective when the posterior parameters are highly correlated.

Implementation of Bayesian models using JAGS

Just Another Gibbs Sampler (JAGS) is a programme for analysis of Bayesian models that uses MCMC methods [Plummer Martyn, 2003]. JAGS uses a Gibbs Sampling approach with several techniques that either enforce or substitute the standard Gibbs Sampler, for example slice sampling [Neal, 2003] or adaptive rejection metropolis sampling [Gilks et al.,

1995]. For this thesis, JAGS was used to implement Bayesian models, along with the `rjags` package in order to work directly with JAGS from within the R language and environment. A JAGS model uses Bayesian inference Using Gibbs Sampling (BUGS) language, in which models are specified using the precision τ as a parameter for the Normal distribution rather than the variance σ^2. The precision is the inverse of the variance, i.e. $\tau = \frac{1}{\sigma^2}$. For example, a Normal distribution with mean 0 and variance 100^2 is identified in BUGS language using `dnorm(0,0.0001)`. Options to consider in the initialisation of a JAGS model include how many parallel chains to run and what initial parameter values should be used (i.e. the starting point for each chain). Using multiple chains with different starting points is recommended in order to determine how well the chains have mixed [Brooks, 1998]. In order to reliably detect convergence to the target (stationary) distribution, initial values should be chosen that are overdispersed with respect to the target distribution [Gelman Andrew and Rubin, 1992]. A random number generator can be used along with a set seed for each chain, in order to make output from a model reproducible. A sampler then acts on the set of parameters in the model, updating these at each iteration. An adaptive phase can be used at the start of the sampling process, whereby samplers used by the model are allowed to change in order to maximise their efficiency. Following this adaptive phase, MCMC output is generally divided into two parts: an initial burn-in period is discarded and the rest of the sample, where the output is considered to have converged sufficiently closely to the target distribution, is used to calculate posterior estimates. The R and JAGS code used to run all models in this chapter (as specified in section 3.5.3) is displayed in Appendix Table A3. An example of a JAGS script used to run a BHM in this chapter is presented in Appendix A4.

"Statistical significance" in Bayesian models

Uncertainty in Bayesian models is quantified via probability distributions, so the reporting of results is in terms of direct probability statements about the parameter values. This is done by calculating the area of the posterior distribution to the right (or left) of a parameter value, which is simply the proportion of values in the posterior sample for that parameter which are greater (or less) than that value. This information is used to report results of Bayesian models as means and 95% posterior *credible* intervals (PCIs) for the parameter estimates. The mean estimate, for example, is the average estimate over the whole of the posterior distribution, and the 95% PCI is the range from the 2.5^{th} to the 97.5^{th} percentile values. In this thesis, if the 95% CI (for frequentist models) or 95% PCI (for Bayesian models) did not include zero then this was considered a "statistically significant"

average annual change in prevalence. Note that, in contrast to individual Poisson models, adjustment for multiple testing was not done for hierarchical models. This is because each group of related CAs are tested together in one model rather than in a number of individual tests, hence when each individual CA is considered, the shrinkage to the mean (across the group of CAs being tested) already provides some level of correction.

Choice of prior distribution

Stipulating prior distributions in a Bayesian analysis is important because, together with the likelihood, these influence the posterior distributions on which we make inferences about parameters of interest. Prior distributions can be defined conveniently using parameters such as means and variances. If an estimate is believed to be accurate then the variance parameter for a prior distribution should be set as a low value and, conversely, where there is a large amount of uncertainty around the estimate a high value should be chosen. There is sometimes prior information available about a parameter of interest, for example through expert opinion or data from previous studies. If such information is available, it should be appropriately summarised by the prior distribution. "Informative" priors have a greater influence on the posterior distribution, which is less dominated by the likelihood when there is more information in the prior. It is often the case, however, that there is no such prior information available, in which case a prior distribution with minimal impact on the posterior should be chosen. These distributions are commonly called non-informative or vague priors, and these enable Bayesian methods to infer estimates for parameters in analyses where there is no further information beyond the data available [Gelman A., 2006]. Non-informative priors are often thought of as reference models, in that they are a starting point in lieu of more informative prior distributions, as considered appropriate for "fully Bayesian" analyses [Bernardo, 1979]. For analyses in this thesis, prior distributions were chosen such that they did not disproportionately affect posterior distributions, generally known as vague or non-informative priors. However, these do contain some information, for example restriction of the prior variance to a certain range. As these priors did provide models with some level of information they are therefore referred to as "minimally" informative. The use of uniform prior distributions for the variance parameters in BHMs has been recommend [Gelman A., 2006], and were used throughout the analyses in this chapter.

Convergence diagnostics for Bayesian models

Parameter inference from the posterior distribution is only valid for MCMC chains that have converged to their stationary (target) distribution. The coda package [Plummer Martyn et

al., 2003] was used to assess model convergence and summarise the sample posterior distribution for each parameter in Bayesian models. Commonly used convergence diagnostics implemented in coda were considered. The Gelman-Rubin diagnostic, for example, aims to identify lack of convergence using multiple parallel chains by calculating the potential scale reduction factor (PSRF) separately for each parameter. The PSRF compares the within and between-chain variances, which should be similar if all chains have achieved convergence to the target distribution. A large PSRF indicates that the between-chain variance is substantially greater than the within-chain variance, meaning longer simulations are required to achieve adequate convergence. Chains are considered "stable" and likely to have reached their target distribution if the PSRF is close to 1. The Heidelberger-Welche and Raftery-Lewis diagnostic tests are used to indicated how long the "burn in" and total sample run (i.e. number of iterations) should be, based on accuracy of the estimation of the mean and quantiles of the posterior distribution, respectively [Plummer Martyn et al., 2003].

Autocorrelation is the correlation of a time series with its own past and future values. This can be problematic in MCMC sampling, where high values of autocorrelation within chains indicate slow mixing and convergence. High autocorrelation means parameter values at successive steps in a chain do not give independent information about the posterior, i.e. the chain is not changing much from one step to the next. It can be useful to "thin" a chain with high autocorrelation before calculating summary statistics, which involves keeping only one value out of every m^{th} step in the chain and discarding all other sampled values. A thinned chain takes up less computing memory and can give a more precise estimate of the posterior sample, but is generally less efficient than using full chain [Link and Eaton, 2012]. It is important to have an idea of how much independent information is contained in each chain and one measure of this is the effective sample size, an estimate of what sample size would be required for a completely non-autocorrelated chain to give the same amount of information [Kruschke, 2014. Section 7.5.2]. The effective sample size is obtained by dividing the actual sample size (i.e. number of iterations in a chain) by the amount of autocorrelation

$$Effective\ Sample\ Size = \frac{N}{1 + 2\sum_{k=1}^{\infty} ACF(k)}$$

Here $\mathrm{ACF(k)}$ is the autocorrelation of the chain at lag k, i.e. two values k steps apart from each other in the chain. An effective sample size of around 10,000 has been recommended as being generally sufficient for accurate and stable estimates of the 95% PCI for parameter values [Kruschke, 2014]. Convergence, mixing of chains and autocorrelation were also

assessed visually. Trace plots were examined for all models parameters to ensure appropriate mixing of chains, given dispersed starting values. Trace plots simply plot the sampled values for each iteration in each chain. Density plots of the posterior distribution for parameters were also used to assess the shape of the posterior distributions.

Based on the above convergence and autocorrelation diagnostic checks, the posterior sample was trimmed and thinned as required, and the posterior distributions were summarised in order to obtain estimates for all parameters of interest.

Evaluation and comparison of models

Evaluation of overall goodness-of-fit and model complexity can be done using measures such as the deviance, a measure of error in which lower values indicate a better fit to the data. The deviance is defined as -2 times the logarithm of the likelihood function [Gelman Andrew and Hill, 2007]

$$D(\theta) = -2\ log\{p(y|\theta)\}$$

Akaike Information Criterion (AIC) is a measure of model fit that includes a penalty based on the number of parameters k in order to discourage overfitting [Akaike, 1973]

$$AIC = -2\ log\{p(y|\hat{\theta})\} + 2k\ =\ D(\hat{\theta}) + 2k$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ. AIC is useful in comparison of non-nested models, and lower AIC values indicate models that give a better fit to the data. The concept of AIC and deviance also apply to hierarchical models, although in this setting the number of parameters is not so clearly defined, since this is dependent on the amount of pooling in the hierarchical model. The conditional Akaike information criterion (cAIC) has been proposed as an adjustment to the AIC, where the random effects structure of multilevel models are taken into account in the estimation of the number of parameters in the model [Vaida and Blanchard, 2005]. Various extensions to the cAIC for GLMMs are described by Saefken et al [2014] and are implemented using the R package `cAIC4`, which uses an analytical estimator to calculate the degrees of freedom [Saefken and Ruegamer, 2014]. The Deviance Information Criterion (DIC) can be thought of as a Bayesian hierarchical modelling equivalent of the cAIC, and is defined as the expected deviance $\overline{D}$ plus the effective number of parameters in the model p_D [Spiegelhalter et al., 2002]

$$DIC = D(\bar{\theta}) + 2p_D = \overline{D} + p_D$$

The effective number of parameters p_D gives an impression of the overall size or complexity of the model, and this quantity depends on the variance of group-level parameters in a model. This is not appropriately captured by AIC because of the presence

of random effects in hierarchical models. In a simple hierarchical model with one group of estimates, for example, the effective number of parameters may be anything from 1 (the average estimate across the group) up to the total number of estimates in the group. The definition of p_D implemented by `rjags` is that proposed by Plummer M [2002]. As with AIC, models with lower DIC are preferable, with p_D favouring models with less parameters in order to compensate for the decrease in DIC due to the value of the expected deviance (i.e. favouring a good fit of the model to the data). The cAIC and DIC were monitored for frequentist and Bayesian models, respectively.

3.5.3. Model specifications

Grouping of CAs according to hierarchies in the EUROCAT subgroup coding were considered, with results from the models outlined in Table 3.1 being compared. GLMMs for model 3 and 4 were implemented in R using the `glmer` command in the `lme4` package [Bates D, 2015]. BHMs were run using JAGS via the R package `rjags`. Examples of R and JAGS code for each model are displayed in Appendix Table A3. A full description and specification of each model is presented in detail in this section and summarised in Appendix Table A4. An example of a script that was used to run a BHM (model 5) for this chapter in R and JAGs is displayed in Appendix A2.

Table 3.1. Summary of models evaluated in Chapter 3 for routine analysis of trends in the prevalence of congenital anomalies (CAs).

Model	Model name	Model description
1	Individual models	Separate Poisson regression models for each CA, pooling over registry
2	Individual models + registry	Separate Poisson regression models for each CA, including a random effect for registry
3	Frequentist hierarchical model	Poisson model for multiple CAs, pooling over registry
4	Frequentist hierarchical model + registry	Poisson model for multiple CAs, including a random effect for registry
5	Bayesian hierarchical model	Bayesian Poisson model for multiple CAs, pooling over registry
6A	Bayesian hierarchical model + registry (A)	Bayesian Poisson model for multiple CAs, including random effect for registry. Random intercepts for registry (across all CAs) and for CA subgroup (across all registries) are included separately
6B	Bayesian hierarchical model + registry (B)	Bayesian Poisson model for multiple CAs, including random effect for registry. Each anomaly-registry combination is allowed a separate random intercept

Model 1: Individual models

Model 1 was a separate Poisson regression for each CA subgroup, pooling over registry to model prevalence rates in years $t = 1, \dots, T$ for a total time period of T years

$$y_t \sim Poisson(\lambda_t)$$

With the log link function

$$\log(\lambda_t) = \log(p_t) + \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t \sim normal(0, \sigma_\varepsilon^2)$$

Here y_t was the number of CA cases for year t in a process with prevalence rate λ_t and total number of births p_t. The coefficient $\beta_1$of x_t was the estimated average annual change in number of cases (on the log scale), and the intercept β_0 represented the expected log prevalence rate in the mean year across all observations due to the predictor variable x_t being centred (i.e. x_t was the difference between the year t and the mean of all the years included in the model $t = 1, \dots, T$). The term ε_t was included in the log link function to account for potential overdispersion (see section 3.5.1). This additional term was included

in practice by the addition of a unique identifier for each observation as a random term in the model.

Model 2: Individual models + registry

Model 2 consisted of separate Poisson regression models for each CA, but additionally included a random effect to account for heterogeneity in the prevalence rates. This was in order to account for heterogeneity in the prevalence rates but not in the trends, which were here assumed to be the same for all registries. This was done for $j = 1, \dots, J$ registries, such that

$$y_{tj} \sim Poisson(\lambda_{tj})$$

$$\log(\lambda_{tj}) = \log(p_{tj}) + (\beta_0 + u_j) + \beta_1 x_{tj} + \varepsilon_{tj}$$

$$u_j \sim normal(0, \sigma_j^2), \qquad \varepsilon_{tj} \sim normal(0, \sigma_\varepsilon^2)$$

In this model, the term u_j denotes the random intercepts for each registry. These were assumed to follow a normal distribution with zero mean and standard deviation σ_j^2. All other terms are as described in model 1.

Model 3: Frequentist hierarchical model

Model 3 was a hierarchical Poisson model that groups similar CAs together, with random effects for each CA and data being pooled over registry. The model for $k = 1, \dots, K$ CAs was defined as

$$y_{tk} \sim Poisson(\lambda_{tk})$$

$$\log(\lambda_{tk}) = \log(p_{tk}) + (\beta_0 + u_{0k}) + (\beta_1 + u_{1k}) x_{tk} + \varepsilon_{tk}$$

$$u_{0k} \sim normal(0, \sigma_{0k}^2), \qquad u_{1k} \sim normal(0, \sigma_{1k}^2), \qquad \varepsilon_{tk} \sim normal(0, \sigma_\varepsilon^2)$$

In model 3 u_{0k} were the random intercepts and u_{1k} the random slopes for each CA, and zero covariance was assumed between u_{0k} and u_{1k}. This implied that the random intercepts and slopes for each anomaly were uncorrelated, i.e. that the prevalence of a CA was not correlated with the trend in prevalence for that CA. The overall intercept and slope in the model were β_0 and β_1 respectively, and the term ε_{tk} allowed for potential overdispersion.

As discussed previously (see section 3.5.2), estimates for the variability around the random effect estimates (for example the random slopes that are the estimates of the average yearly prevalence for each anomaly subgroup) are not reported using frequentist GLMMs, therefore CIs were not calculated for the intercepts and slopes for each anomaly subgroup in this model.

Model 4: Frequentist hierarchical model + registry

Model 4 was a hierarchical Poisson model grouping CAs, with random effects for each CA and an additional random intercept for registry. This was defined for $k = 1, \dots, K$ CAs and $j = 1, \dots, J$ registries, such that

$$y_{tjk} \sim Poisson(\lambda_{tjk})$$

$$\log(\lambda_{tjk}) = \log(p_{tjk}) + (\beta_0 + u_{0k} + u_j) + (\beta_1 + u_{1k})x_{tjk} + \varepsilon_{tjk}$$

$$u_{0k} \sim normal(0, \sigma_{0k}^2), \qquad u_{1k} \sim normal(0, \sigma_{1k}^2), \qquad u_j \sim normal(0, \sigma_j^2),$$

$$\varepsilon_{tjk} \sim normal(0, \sigma_\varepsilon^2)$$

Random intercepts for registry were again denoted by u_j, random intercepts and slopes for each CA were u_{0k} and u_{1k}, respectively, and zero covariance between the random effect parameters was assumed. As in model 3, the mean intercept and slope were β_0 and β_1, respectively, and ε_{tk} accounted for potential overdispersion in the data. Again, as in model 3, CIs were not calculated for individual intercepts and slopes in this model since estimates of the variability around the random effects are not reported.

Model 5: Bayesian hierarchical model

Model 5 was a Bayesian specification of model 3: a Poisson BHM including multiple CAs and pooling data over registry. In this model, 95% PCIs were estimated for the random effects for $k = 1, \dots, K$ CAs using the 2.5th and 97.5th percentile values of the posterior distribution for each parameter. The model is specified as follows

Likelihood model:

$$y_{tk} \sim Poisson(\lambda_{tk})$$

$$\log(\lambda_{tk}) = \log(p_{tk}) + u_{0k} + u_{1k}x_{tk} + \varepsilon_{tk}$$

Priors for model parameters:

$$u_{0k} \sim normal(\mu_{u0}, \tau_{u0}), \qquad u_{1k} \sim normal(\mu_{u1}, \tau_{u1}),$$

$$\varepsilon_{tk} \sim normal(0, \tau_\varepsilon)$$

Hyper-priors for model parameters:

$$\mu_{u0},\ \mu_{u1} \sim normal(0, 0.001)$$

$$\tau_{u0} = \frac{1}{\sigma_{u0}^2}, \qquad \tau_{u1} = \frac{1}{\sigma_{u1}^2}, \qquad \tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2},$$

$$\sigma_{u0}, \sigma_{u1}, \sigma_{\varepsilon k} \sim uniform(0, 10)$$

In this model, μ_{u0} denoted the mean intercept across the anomaly subgroups and μ_{u1} the mean trend. As in model 3, the random intercepts and slopes for each anomaly subgroup were denoted by u_{0k} and u_{1k}, respectively, and zero covariance was assumed between random effect parameters. Minimally informative prior distributions were used for all parameters, with parameters for prior distributions as specified above.

Model 6A: Bayesian hierarchical model + registry (A)

Model 6A was a Poisson BHM including multiple CAs and a random registry effect. This was a Bayesian equivalent of model 4, which allowed estimation of the variability around the random effects.

Likelihood model:

$$y_{tjk} \sim Poisson(\lambda_{tjk})$$

$$\log(\lambda_{tjk}) = \log(p_{tjk}) + u_j + u_{0k} + u_{1k} x_{tjk} + \varepsilon_{tjk}$$

Priors for model parameters:

$$u_{0k} \sim normal(\mu_{u0}, \tau_{u0}), \qquad u_{1k} \sim normal(\mu_{u1}, \tau_{u1}),$$

$$u_j \sim normal(\mu_j, \tau_j)$$

$$\varepsilon_{tjk} \sim normal(0, \tau_\varepsilon)$$

Hyper-priors for model parameters:

$$\mu_{u0},\ \mu_{u1} \sim normal(0, 0.001)$$

$$\tau_{u0} = \frac{1}{\sigma_{u0}^2}, \qquad \tau_{u1} = \frac{1}{\sigma_{u1}^2}, \qquad \tau_j = \frac{1}{\sigma_j^2}, \qquad \tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2},$$

$$\sigma_{u0}, \sigma_{u1}, \sigma_j, \sigma_\varepsilon \sim uniform(0, 10)$$

In model 6A, u_j denoted the random intercepts for registry and all other estimates were as described for model 5.

Model 6B: Bayesian hierarchical model + registry (B)

Model 6B was a Poisson BHM including multiple CAs and a random registry effect, differing from model 6A in that it each CA-registry combination was allowed to have a separate random intercept.

Likelihood model:

$$y_{tjk} \sim Poisson(\lambda_{tjk})$$

$$\log(\lambda_{tjk}) = \log(p_{tjk}) + u_{0jk} + u_{1k}x_{tjk} + \varepsilon_{tjk}$$

Priors for model parameters:

$$u_{0jk} \sim normal(\mu_{u0}, \tau_{u0}), \qquad u_{1k} \sim normal(\mu_{u1}, \tau_{u1}),$$

$$\varepsilon_{tk} \sim normal(0, \tau_{\varepsilon})$$

Hyper-priors for model parameters:

$$\mu_{u0},\ \mu_{u1} \sim normal(0, 0.001)$$

$$\tau_{u0} = \frac{1}{\sigma_{u0}^2}, \qquad \tau_{u1} = \frac{1}{\sigma_{u1}^2}, \qquad \tau_{\varepsilon} = \frac{1}{\sigma_{\varepsilon}^2},$$

$$\sigma_{u0}, \sigma_{u1}, \sigma_{\varepsilon} \sim uniform(0, 10)$$

Here u_{0jk} represented JxK random intercepts (i.e. one for each anomaly-registry combination), and other parameters were as specified in model 6A.

3.5.4. Sensitivity analyses

Estimating trends separately in each registry

The prevalence of many CAs is known to vary considerably between different regions in Europe and this is usually due to differences in reporting (e.g. due to availability of prenatal ultrasound in a particular population), although it may potentially be the case that an environmental risk factor is present only in a particular population. However, any general trend in prevalence may be expected to be broadly similar across the registries. Therefore, in the analysis of NTDs, the average annual trend in the prevalence of the subgroups was also estimated using a separate Poisson regression model for each registry-anomaly combination. This aimed to identify whether there were registries with trends markedly inconsistent with the overall European trend, and might have disproportionate influence on results in models that included all registries together.

Use of alternative samplers in Bayesian models

Where there was evidence of strong autocorrelation in Bayesian models, the use of alternative samplers were investigated using the programme Stan [Stan Development Team, 2015b]. This was done to assess whether the alternative samplers used by Stan (a Hamiltonian Monte Carlo variant called the No U-Turn sampler) might provide a solution in the case where the Gibbs samplers do not converge due to high autocorrelation. Models

were fitted in Stan via R, using the RStan package [Stan Development Team, 2015a] with the same model parameters as those described for JAGS models.

Use of alternative prior distributions in Bayesian models

Different choices of parameters for the prior distributions used in Bayesian models were considered, as discussed in section 3.5.2. When the number of groups in a BHM is small (i.e. less than 5), posterior distributions for the variance parameters tend to have long right-tails. This can lead to unrealistically high values for estimated standard deviation parameters for the random effect and result in an "under shrinking" of the estimates of the random effects [Gelman A., 2006]. This means that the trends for each subgroup in a BHM may not demonstrate as much shrinkage as expected if there are only a small number of CAs in the model. In such situations, the half-Cauchy prior distribution with a weak constraint on the standard deviation has been shown to perform better than a uniform distribution [Gelman A., 2006]. A weakly informative half-Cauchy prior in this context is intended to constrain the posterior distribution, rather than being a representation of our actual state of prior knowledge. As a sensitivity analysis, therefore, a half-Cauchy prior was also evaluated for the variance parameters in models with less than five CAs. Appendix A3 contains details about how this half-Cauchy prior distribution was implemented in JAGS.

3.6. Results

3.6.1. Description of EUROCAT prevalence data in 18 registries

Data was available from 18 EUROCAT registries in 11 countries, with each registry having available data for at least nine years of the period 2003 to 2012 as required by the study inclusion criteria. The earliest year of data available ranged from 1983 (in six registries) to 2002 (in Isle da la Reunion registry). The data were aggregated according to registry, year of birth, sex and maternal age (in five-year age groups), including and excluding chromosomal cases separately. For aggregate counts of non-chromosomal CA cases, the following exclusions were made: chromosomal anomalies, genetic syndromes or microdeletions, teratogenic syndromes with malformations, and sequences. Maternal age was defined as age at delivery of baby, and this was known for 96.2% of the population included in this study. Within each registry and year, the remaining 3.8% were assumed to follow the same maternal age distribution as those with known maternal age.

The total prevalence of all CAs for the period 2003 to 2012 is displayed in Table 3.2, including and excluding chromosomal subgroups separately. Over all registries, the combined total of births in the population covered by these registries for these years of data was just over 4 million. This included 103,507 (2.5%) cases of CA, 81,147 of which did not have a chromosomal anomaly (78.4% of all cases). The highest total prevalence of all anomalies was in Vaud (3.8%) and the lowest was in Dublin (1.7%).

Overall, there was a higher number of male CA cases (54.7%) than females (42.1%), with the remainder having indeterminate sex (0.1%) or listed as “Not known, or missing” (3.2%). The majority of mothers (71.4%) were under the age of 35 at the time of delivery, with only 5% of mothers being under 20 and 0.5% of mothers over 45 years old.

Table 3.2. Total number of births, cases and prevalence of congenital anomalies per 100 births in 18 EUROCAT registries from 2003-2012, including and excluding genetic conditions.

			Total		**Total excluding genetic conditions**	
Country	Registry	Total Births 2003-2012 [a]	Cases	Prevalence per 100 births (95% CI)	Cases	Prevalence per 100 births (95% CI)
Austria	Styria	103,492	3,193	3.09 (2.98, 3.19)	2,675	2.58 (2.49, 2.68)
Belgium	Antwerp	200,819	4,839	2.41 (2.34, 2.48)	3,991	1.99 (1.93, 2.05)
	Hainaut	126,689	3,021	2.38 (2.30, 2.47)	2,406	1.90 (1.82, 1.98)
Denmark	Odense	51,693	1,532	2.96 (2.82, 3.12)	1,194	2.31 (2.18, 2.44)
France	Paris	266,387	8,536	3.20 (3.14, 3.27)	6,302	2.37 (2.31, 2.42)
	Isle de la Reunion	146,462	3,974	2.71 (2.63, 2.80)	3,115	2.13 (2.05, 2.2)
Germany	Saxony-Anhalt	172,272	5,526	3.21 (3.12, 3.29)	4,876	2.83 (2.75, 2.91)
Ireland	Cork and Kerry	96,833	2,580	2.66 (2.56, 2.77)	2,024	2.09 (2.00, 2.18)
	Dublin	256,377	4,345	1.69 (1.64, 1.75)	3,172	1.24 (1.19, 1.28)
Italy	Tuscany	299,863	6,335	2.11 (2.06, 2.17)	5,234	1.75 (1.70, 1.79)
Netherlands	Northern Netherlands	180,927	4,832	2.67 (2.60, 2.75)	3,891	2.15 (2.08, 2.22)
Spain	Basque Country	184,570	4,503	2.44 (2.37, 2.51)	3,335	1.81 (1.75, 1.87)
Switzerland	Vaud	76,241	2,881	3.78 (3.64, 3.92)	2,186	2.87 (2.75, 2.99)
UK	East Midlands & South Yorkshire	717,264	15,335	2.14 (2.10, 2.17)	12,294	1.71 (1.68, 1.74)
	Northern England	328,496	7,843	2.39 (2.34, 2.44)	5,982	1.82 (1.78, 1.87)
	Thames Valley	254,090	5,224	2.06 (2.00, 2.11)	3,677	1.45 (1.40, 1.49)
	Wales	343,245	12,668	3.69 (3.63, 3.76)	10,567	3.08 (3.02, 3.14)
	Wessex	291,422	6,340	2.18 (2.12, 2.23)	4,226	1.45 (1.41, 1.49)
All registries combined		**4,097,142**	**103,507**	**2.53 (2.51, 2.54)**	**81,147**	**1.98 (1.97, 1.99)**

[a] Basque country did not include data for 2012

3.6.2. Accounting for potential overdispersion: encephalocele as an example

Counts of encephalocele per year for the latest ten years of data are displayed in Figure 3.3. A number of the registries have a high number of years with no cases of encephalocele, i.e. zero counts. Over 40% of the yearly counts across all 18 registries were zero, however when counts were combined for all registries the smallest number of total cases in one year was 29.

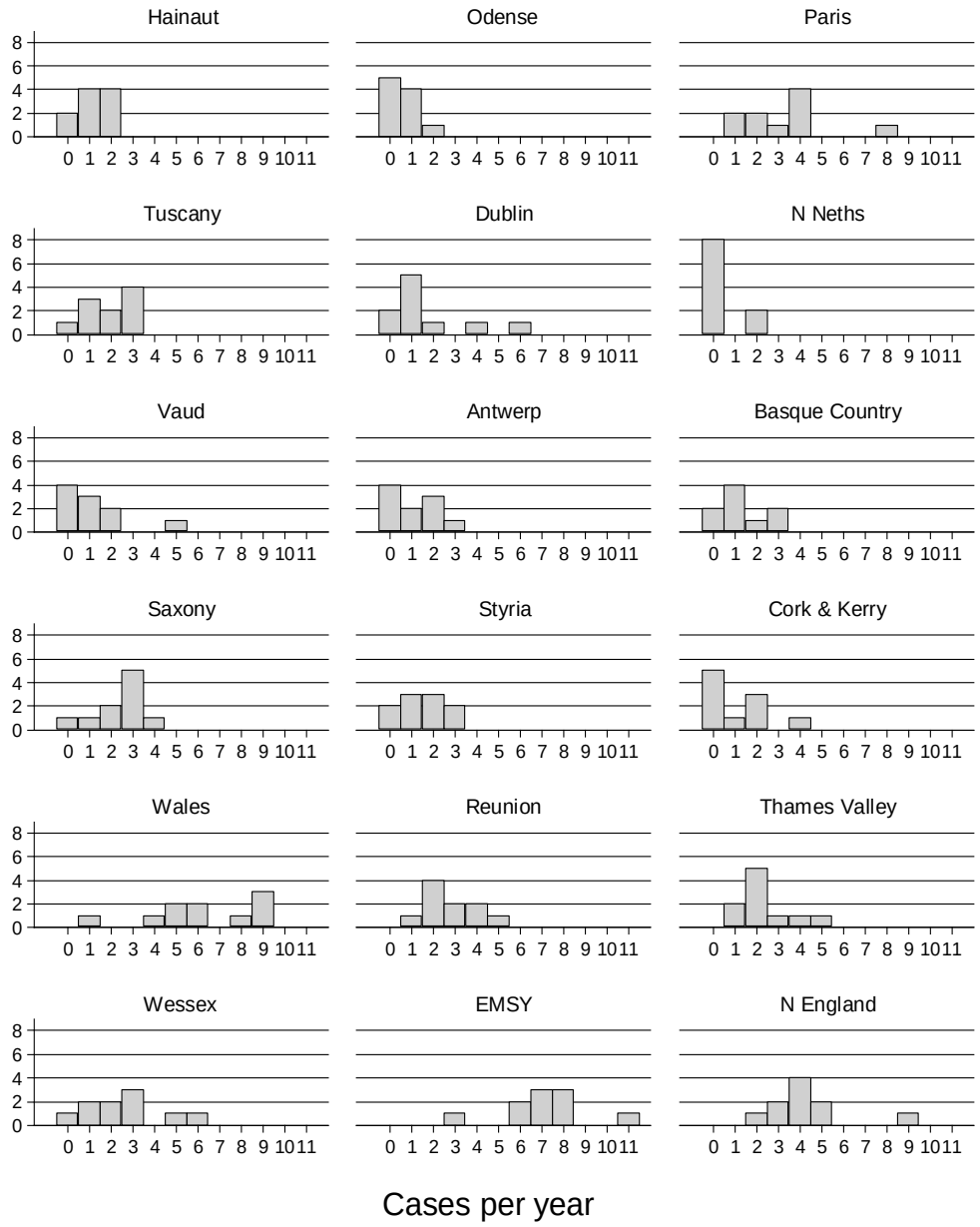


Figure 3.3. Yearly counts of encephalocele cases in 18 EUROCAT registries from 2003-2012.

Figure 3.4 plots the mean against the variance of the encephalocele counts per year, showing that the mean and variance are similar for some registries, e.g. Odense and Antwerp, where the markers are close to the diagonal line of mean=variance. Points

noticeably above the diagonal line indicate overdispersion in the distribution of yearly counts for those registries, and those under the line indicate some amount of under-dispersion. The majority of the points in Figure 3.4 lie fairly close to the diagonal line, indicating that there is not clear under or overdispersion for the encephalocele prevalence data.

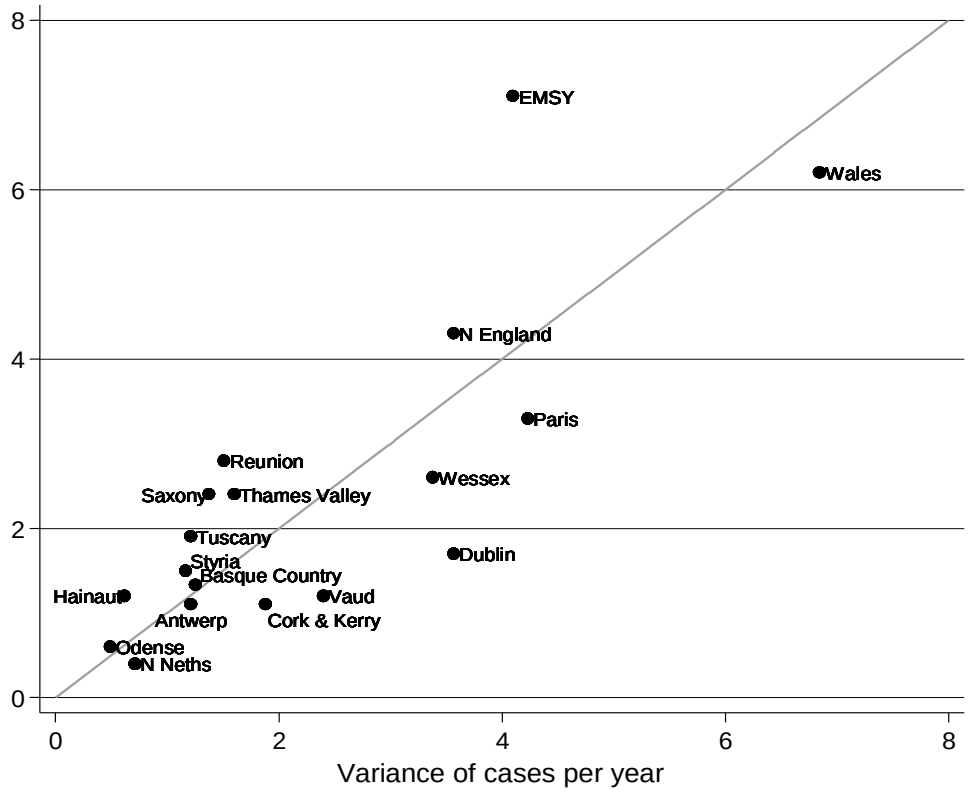


Figure 3.4. Mean against variance of yearly counts of encephalocele in 18 EUROCAT registries from 2003-2012.

Results from different models taking departures from equidispersion into account when assessing trends in encephalocele are displayed in Table 3.3, including different distributions as well as different ways of including the registry effect in the model. All models estimated the same average yearly decrease in encephalocele, which was not a significant trend in any model. Furthermore, there was no significant evidence of dispersion for the data in any of the models considered here, with the dispersion parameter being estimated as not significantly different from 1 (where 1 indicates equidispersion) in all models. The estimated value of the shape parameter of the negative binomial distribution was very large in both the negative binomial models, with the iteration limit being reached whilst fitting this parameter (i.e. the convergence for the value of theta was not achieved). In the Poisson-lognormal model for encephalocele, the standard deviation of the additional

overdispersion term was estimated to be zero. The different models considered for encephalocele were also assessed for a number of other rare CAs (including cystic adenomatous malformation of the lung and anotia). There was no indication for any of the CAs considered that there was strong overdispersion in the data or that the choice of model used to account for the overdispersion was having any effect on the estimated trends in prevalence.

Based on these results, it was decided that a Poisson-lognormal model would be used for all further models considered. The standard deviation of this additional term was monitored throughout all models.

Table 3.3. Summary of results from seven different models considering potential overdispersion in the yearly counts of encephalocele cases in 18 EUROCAT registries.

Model	Treatment of registry in model	Estimated average yearly trend (95% CI)	Dispersion parameter [a]	Degrees of freedom	P-value for dispersion
Poisson GLM	Pooled	-0.021 (-0.05, 0.01)	0.863	8	0.547
Poisson GLM	Adjusted	-0.022 (-0.06, 0.01)	1.075	160	0.246
Quasi-Poisson GLM	Adjusted	-0.022 (-0.06, 0.01)	n/a*	160	n/a
Negative binomial GLM	Adjusted	-0.022 (-0.06, 0.01)	1.074	160	0.246
Poisson GLMM	Random effect	-0.021 (-0.05, 0.01)	0.977	176	0.572
Negative binomial GLMM	Random effect	-0.021 (-0.05, 0.01)	0.977	176	0.572
Poisson-lognormal GLMM	Random effect	-0.021 (-0.05, 0.01)	0.983	175	0.551

[a] The dispersion parameter for the quasi-Poisson model is the same as estimated in the Poisson model

A note on the presentation of results in chapter 3

Full results from all models in all CAs considered in this chapter are displayed throughout Appendix A4 – A7. Note that models 3 and 4 estimated trends for each CA as random effects in a frequentist setting and, as discussed previously, estimates of the variability around these parameters were not calculated (see section 3.5). Results for models 3 and 4 therefore are not included in figures that show estimated trends in prevalence for groups of CAs throughout this chapter. However, point estimates obtained using these models (3 or 4) were in practice very close to those obtained when using the equivalent BHMs (i.e. models 5, 6A and 6B).

3.6.3. Neural tube defects

Figure 3.5 shows estimates of the total prevalence and 95% CIs for each of the 18 EUROCAT registries from 2003-2012 in the three NTD subgroups and for the overall NTD subgroup. The European average prevalence with its 99% confidence intervals is shown as a vertical grey shaded band in order to visualise registries whose NTD prevalence was inconsistent with the European average. Several registries showed prevalence markedly different to the European-wide prevalence for each of the NTD subgroups, and in particular for the overall NTD prevalence. For anencephaly, for example, 11 of the 18 registries had 95% confidence limits that did not overlap with the 99% confidence limits for the European average. The rarest NTD was encephalocele, which had an average total prevalence of just above 1 case per 10,000 births across Europe over the last ten years of data. Prevalence of anencephaly and spina bifida was around four and five cases per 10,000 births, respectively. The total prevalence of all the NTDs combined across the 18 registries was around 10-11 cases per 10,000 births.

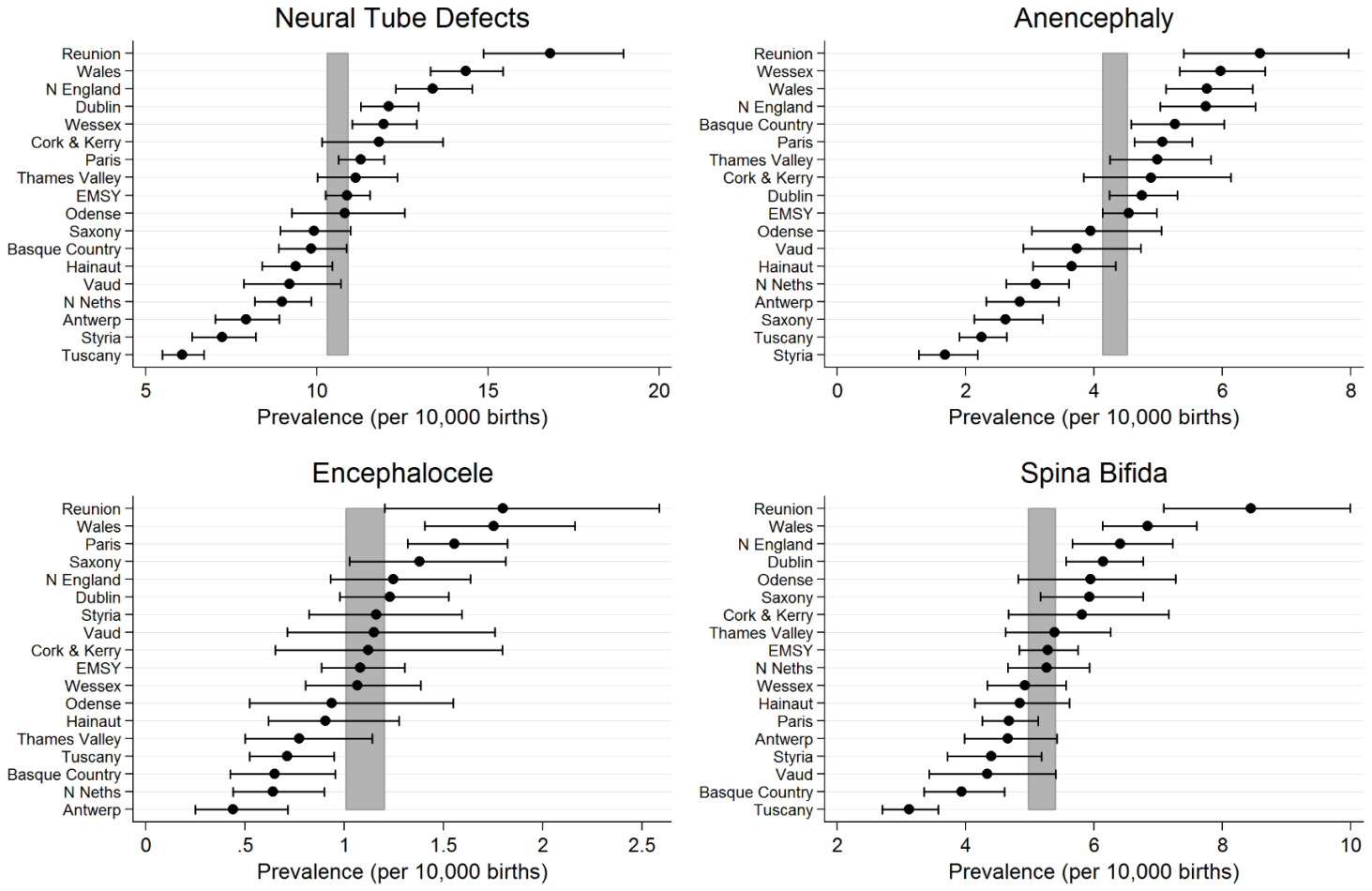


Figure 3.5. Total prevalence of neural tube defects and 95% confidence intervals in 18 EUROCAT registries from 2003 to 2012, with 99% confidence range for the average prevalence across all registries marked as grey shaded bands for each anomaly.

The yearly prevalence and 95% confidence intervals across all 18 registries is displayed in Figure 3.6, showing some fluctuations in NTD prevalence across this period, but no clear trends or patterns for any NTD subgroups.

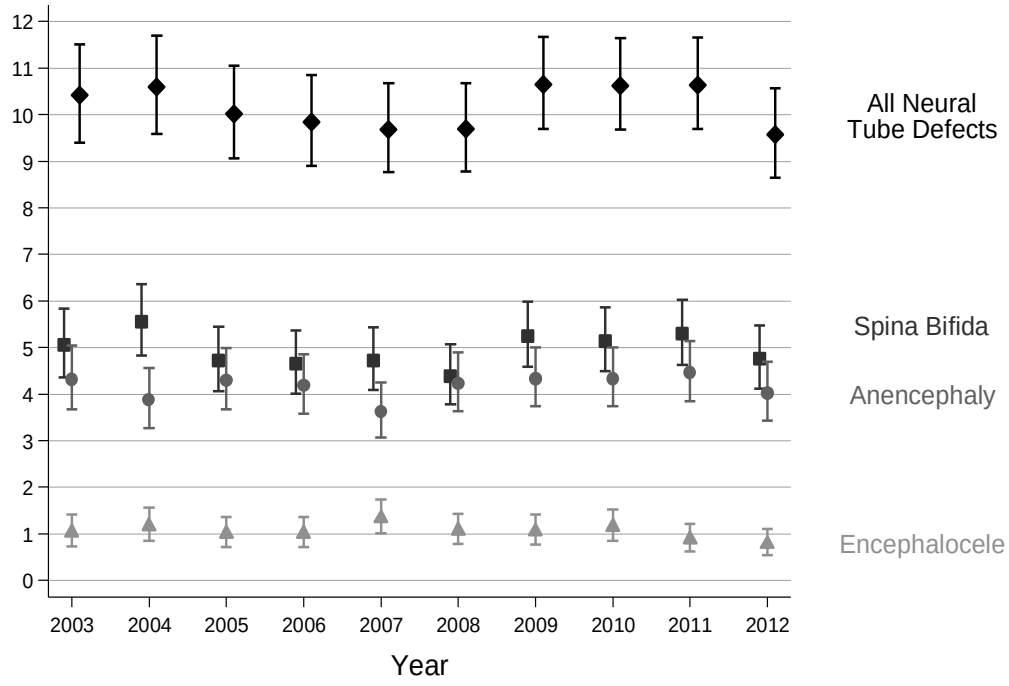


Figure 3.6. Average yearly prevalence and 95% confidence intervals for neural tube defects across 18 EUROCAT registries from 2003-2012.

Estimated average annual trends in the prevalence of NTDs when considered separately (individual models) and combined (hierarchical models) as specified in section 3.5.3 are shown in Figure 3.7. For the aggregate group of the three NTDs combined, estimates from individual models are the total prevalence and 95% CI. For hierarchical models, the average of the estimated slopes for the three CAs are displayed, i.e. the 95% PCI for the parameter μ_{u0} (see section 3.5.3). Figure 3.7 shows that results across all models were consistent, with no evidence of trends in prevalence for any of the NTD subgroups whether considering CAs separately or combining them together in a hierarchical model. The average trend across the NTD subgroups was slightly decreasing, although this effect had a very large variance. In hierarchical models grouping the three NTDs together, there was a small amount of shrinkage towards this average trend. This can be seen in the estimates for encephalocele, which are slightly higher for BHMs than in individual models, and in those for spina bifida and anencephaly, which are very slightly lower. Appendix Table A5 includes further details regarding estimates from models 1-4 in NTD subgroups.

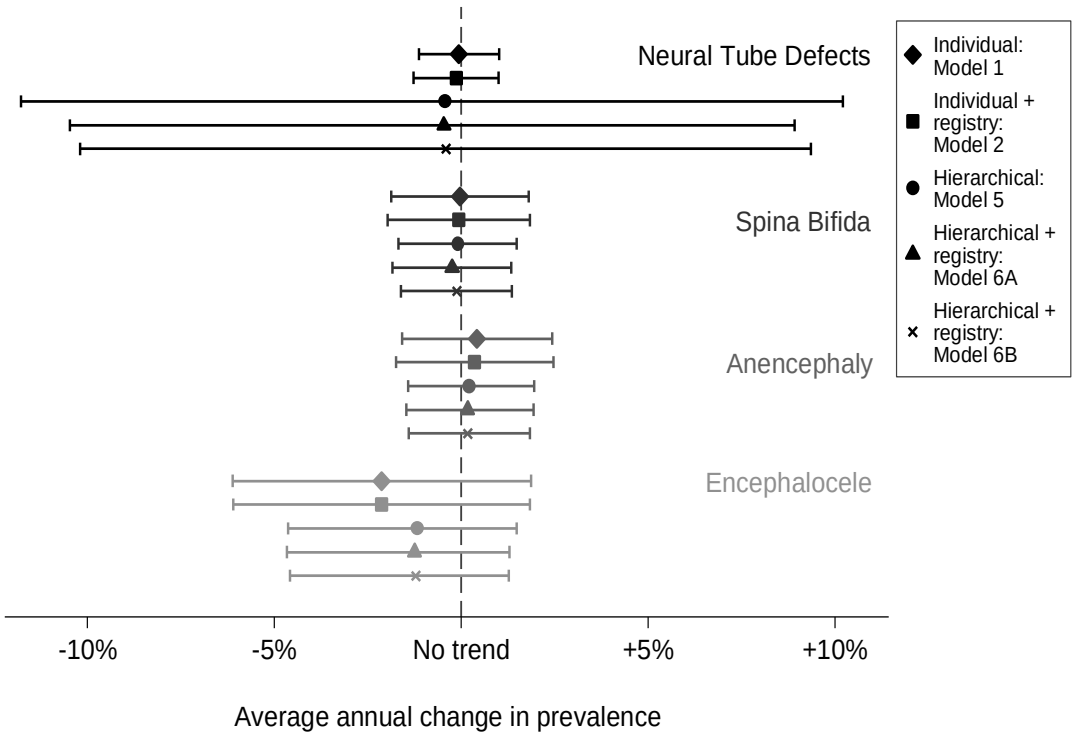


Figure 3.7. Average annual trends in prevalence of neural tube defects; estimates and 95% confidence intervals from individual and hierarchical models as described in section 3.5.3.

Assessment of hierarchical models for neural tube defects

Convergence and model fit were assessed both graphically and using tests as described previously (section 3.5.2). Features of model fit for all the hierarchical models are displayed in Table 3.4. In frequentist hierarchical models 3 and 4, the CA intercepts and slopes were perfectly correlated (correlation coefficient r=1). These high levels of correlation are likely due to over-parameterisation, demonstrating that the models are attempting to estimate more parameters than the data can support. This means that both random effects (i.e. the intercepts and the slope) are not able to be reliably estimated. This is reflected in the small variance in the random effects; in models 3 and 4, there was very little difference in the estimated trends for the three NTDs and the estimated SD of the trends in the three NTDs was around 0.01. This is considerably smaller than the estimated SD of the intercepts (around 0.7). Over-parameterisation in the model was also was evident in the high autocorrelation and poor convergence of the estimated trend parameters as well as in the means and SDs for the trend parameters in the BHMs (models 5, 6A and 6B; see Appendix figures A5, A7-8, A10-14, A18).

The estimated standard deviation of the mean trend across the subgroups was small in all models, indicating little variability amongst the yearly trends in each NTD subgroup, i.e. the

slope for each CA was similar (and in this case non-significant). The cAIC and Deviance were larger for models including registry as a random effect compared to those pooling over registry, reflecting the increased number of parameters added to the model when including a registry effect. The estimated standard deviation of the dispersion parameter was similar for all models, and some overdispersion was present for all models. Model notation for BHMs grouping NTD subgroups is summarised in Appendix Table A6.

In the BHM that pooled information across registries (model 5), there was generally good mixing of chains and low levels of autocorrelation for the majority of parameters. Based on 3 separate chains, the Raftery-Lewis diagnostic for this model indicated that a total sample size of around 14,100 iterations would be required to estimate the 95% PCI for all parameters in the model with an actual posterior probability between 92.5% and 97.5%.

Table 3.4. Model fit in hierarchical models for analysis of trends in the prevalence of neural tube defects.

	***3**: Frequentist hierarchical model*	***4**: Frequentist hierarchical model + registry*	***5**: Bayesian hierarchical model*	***6A**: Bayesian hierarchical model + registry (A)*	***6B**: Bayesian hierarchical model + registry (B)*
cAIC	232	2466	-	-	-
Deviance	242	2513	-	-	-
Residual DF	24	530	-	-	-
Mean Deviance	-	-	223	2395	2362
Penalty	-	-	9	72	68
DIC	-	-	231	2467	2430
Multivariate PSRF	-	-	1.02	3.47	1.02
Mean SD for overdispersion parameter	0	0.124	0.029	0.120	0.064
Mean SD of registry intercepts	-	0.296	-	0.322	2.268 [a]
Mean SD of CA intercepts	0.692	0.695	2.282	2.311	
Mean SD of CA trends	0.010	0.009	0.070	0.054	0.059
Estimated correlation between CA intercepts and slopes	1	1	-	-	-

[a] *SD of the intercepts across all registry-CA combinations*

Model parameters that averaged across the subgroups required much larger posterior samples, with parameters for each NTD subgroup requiring around 1,100 iterations according to the Raftery-Lewis diagnostic (see Appendix A5 and Table A7 for details). For each BHM, therefore, 3 separate chains of 100,000 iterations with a thin of 5 were used, resulting in a posterior sample size of 20,000 iterations per chain. The parameter for the estimated variability of the mean of the trends for the NTD subgroups `sigma.u1` in model 5 showed high levels of autocorrelation for all three chains, and a low effective sample size compared to other parameters in the model (Appendix Table A8). The effective sample size for the overdispersion parameter `sigma.e` in model 5 was also relatively low and with high levels of autocorrelation, in particular for one of the three chains. These two parameters also had estimated dependence factors well above 5 following the Raftery-Lewis diagnostics (Appendix Table A7). Figure 3.8 shows the trace, density and autocorrelation plot for the estimated annual yearly European trend in spina bifida from model 5 (parameter `u1[3]`), as an example of a parameter with good mixing of chains, convergence and low autocorrelation. Each chain is displayed in a different colour (pink, green or blue).

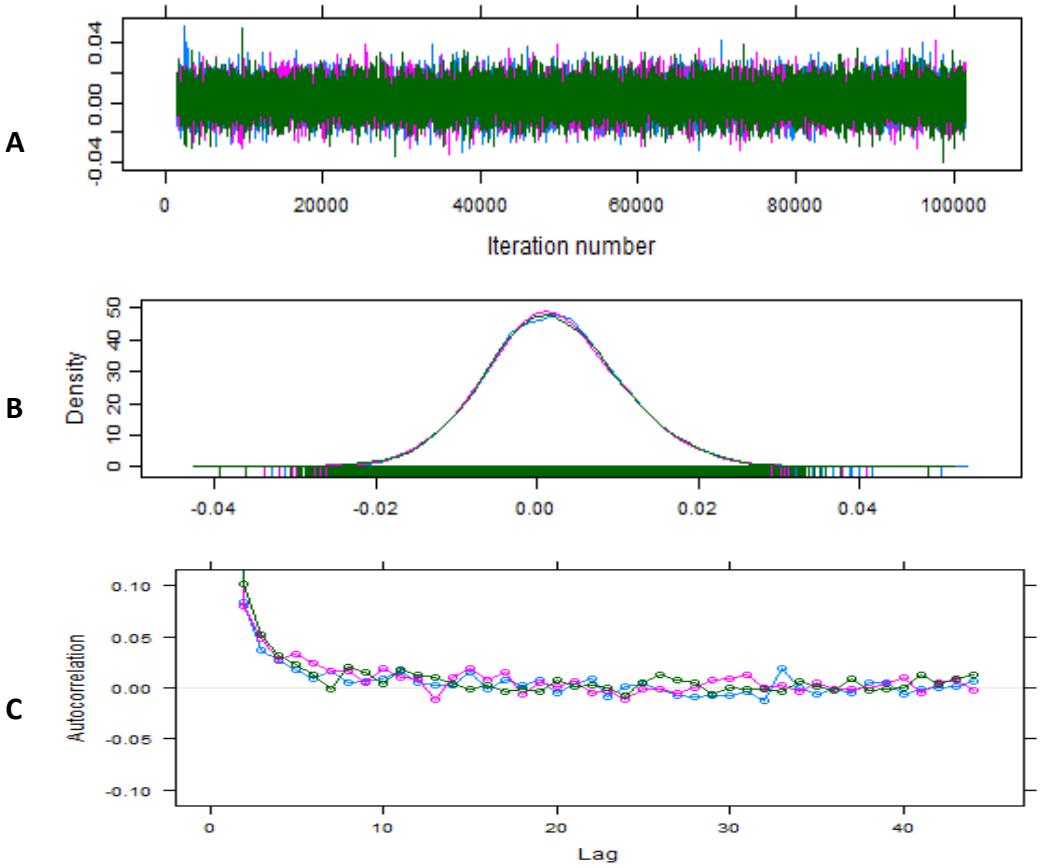


Figure 3.8. Example of a trace (**A**), density (**B**) and autocorrelation (**C**) plot for a parameter with good convergence and mixing of chains and low autocorrelation.

The trace (Figure A1 and A2), density (Figure A3 and A4) and autocorrelation (Figure A5 and A6) plots for all parameters in model 5 are displayed in Appendix A4.

Raftery-Lewis diagnostics, summary of the posterior distribution and trace, density and autocorrelation plots for models 6A (Table A9-A10 and Figures A7-A13) and 6B (Table A11-A12 and Figures A14-A21) are displayed in Appendix A5. In both models 6A and 6B, the parameters $u1[k]$ for the trend in each NTD subgroup showed good mixing and convergence and high effective sample sizes. However, there was high dependency, poor mixing of chains, a lack of convergence and very high autocorrelation for the random intercepts parameters for registry and CA in model 6A. Effective sample sizes for these parameters were extremely low, for example, the random intercepts had an effective sample size of only ~12 for a total sample of 20,000 iterations per chain (Appendix Table A9). In model 6B, effective sample sizes were high for all parameters except for sigma.u1, the estimated variability around the mean of the trends for the NTD subgroups, and `sigma.e`, the overdispersion parameter, which both showed poor mixing of chains, high dependency and high levels of autocorrelation (Appendix Table A10 and A11). The estimated multivariate PSRF from the Gelman and Rubin diagnostic test was close to 1 for models 5 and 6B, but was substantially greater than 1 for model 6A, indicating an overall lack of convergence to the stationary distribution (Table 3.4).

Figure 3.9 shows the trace, density and autocorrelation plot for the intercept of the trend in spina bifida from model 6A, as an example of a parameter with poor mixing of chains, lack of convergence and very high autocorrelation.

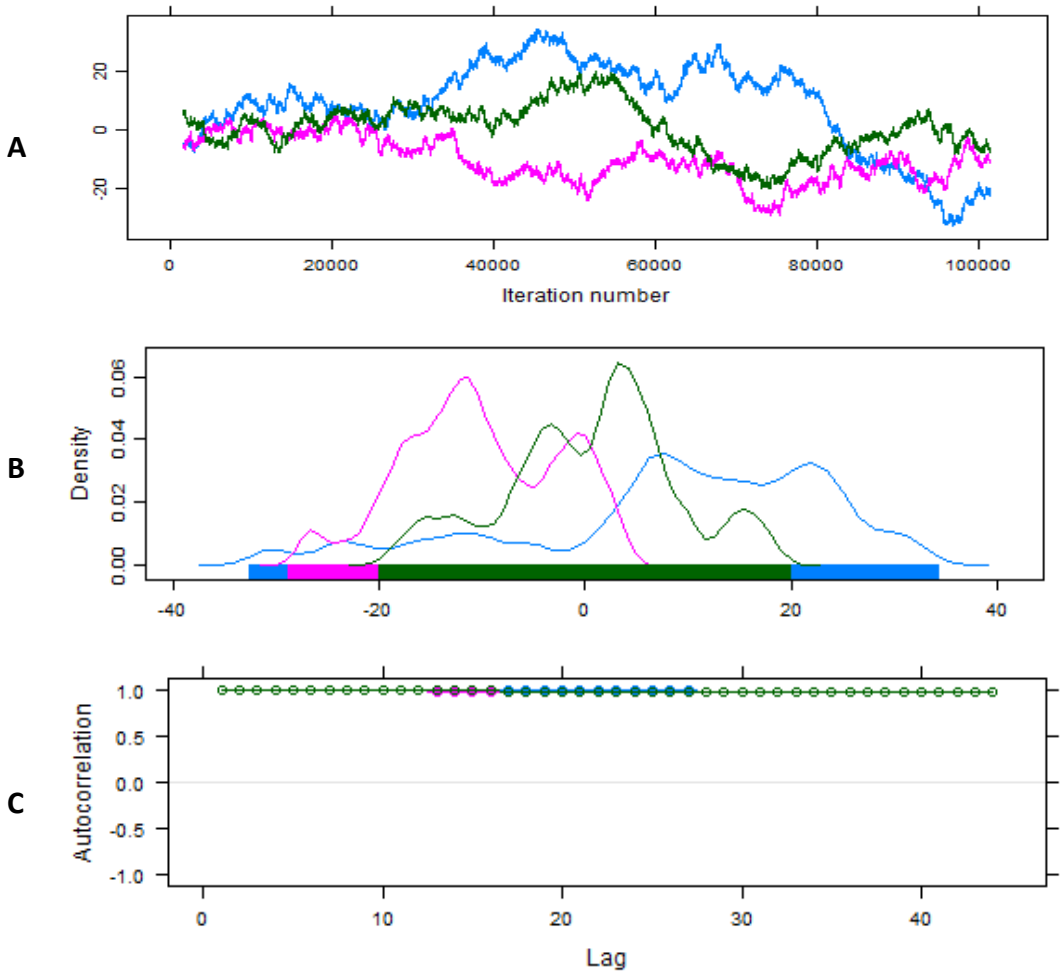


Figure 3.9. Example of a trace (**A**), density (**B**) and autocorrelation (**C**) plot for a parameter with lack of convergence, poor mixing of chains and very high autocorrelation.

Sensitivity analysis I: individual models for each registry

The average annual trend in prevalence of NTD subgroups for 2003-2012 was also estimated using a separate Poisson regression model for each registry-CA combination. The trend in each of the NTD subgroups was consistent across all registries, with estimates and 99% confidence limits for all registries overlapping or being very close to the 99% confidence band for the average European trend (Appendix Figure A21).

Sensitivity analysis II: use of alternative sampling for MCMC chains

Parameters in models that grouped CAs together were highly correlated. Since the Gibbs samplers used by JAGS can be ineffective in such a setting, the programme Stan was therefore also used to see whether its alternative samplers would provide a solution where

the Gibbs samplers did not converge. Models 5 and 6 were fitted using Stan with the same model parameters as those used in JAGS models, i.e. the same values for prior distributions, burn-in, thinning and total number of iterations.

Table 3.5 shows that estimates of the average yearly trend in prevalence for NTDs from models using JAGS Gibbs sampling and Stan's NUTS sampler were very similar. There was some difference in the estimated overall mean trend across the three NTDs in model 5, but estimates were nonsignificant in both cases. The estimated standard deviation of the NTD trends was slightly larger in Stan for model 6A.

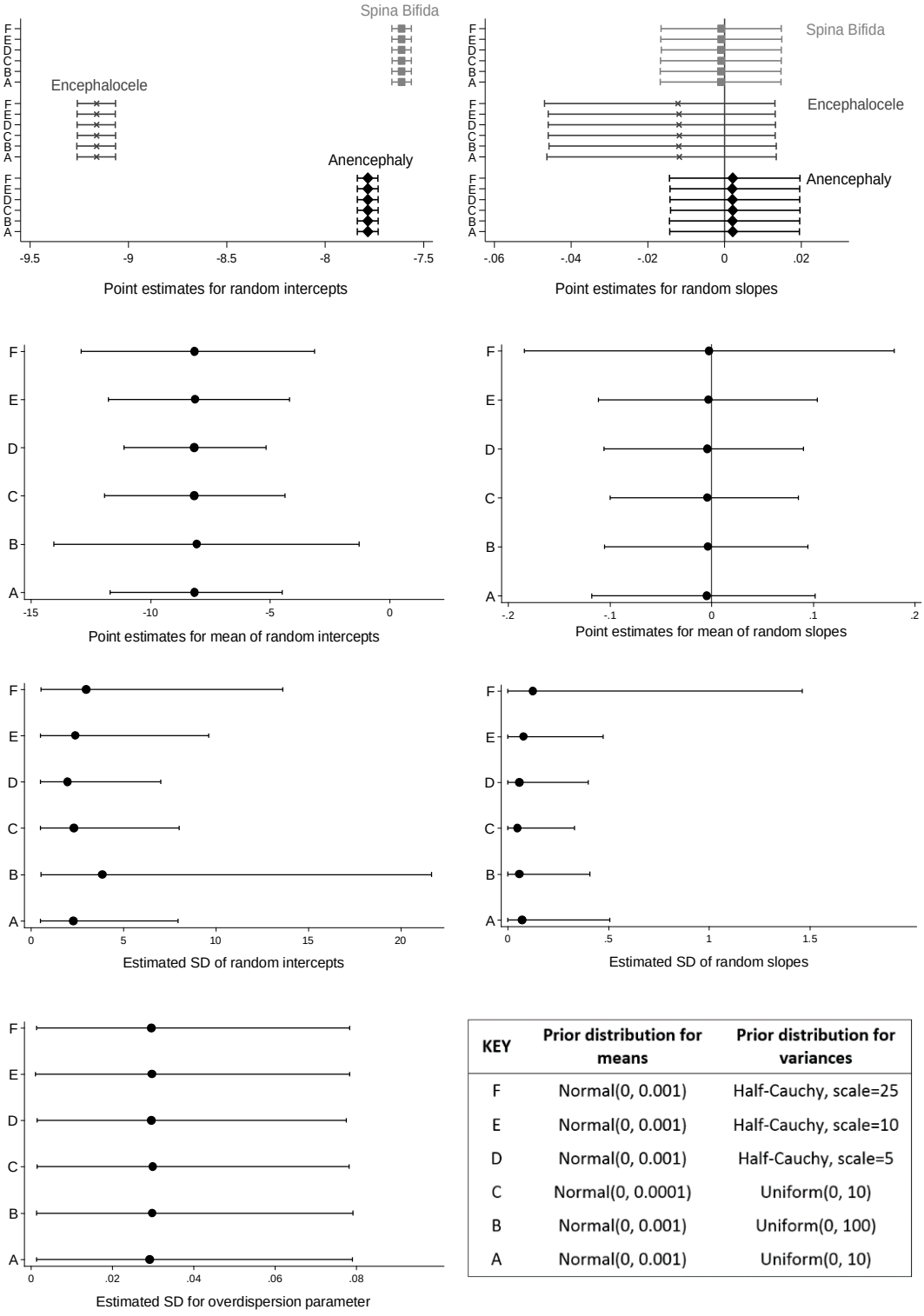
Table 3.5. Comparison of estimated trends in hierarchical models for neural tube defects using JAGS and Stan.

Model	Congenital Anomaly	JAGS	Stan	Ratio of point estimates
5	Anencephaly	0.002 (-0.014, 0.020)	0.001 (-0.013, 0.019)	3.0
	Encephalocele	-0.012 (-0.046, 0.013)	-0.012 (-0.046, 0.012)	1.0
	Spina bifida	-0.001 (-0.017, 0.015)	-0.001 (-0.016, 0.014)	1.8
	Mean of trends for NTD subgroups	-0.004 (-0.118, 0.102)	-0.027 (-0.147, 0.074)	0.2
	SD of trends for NTD subgroups	0.070 (0.001, 0.508)	0.083 (0.001, 0.470)	0.8
6A	Anencephaly	0.002 (-0.015, 0.019)	0.002 (-0.014, 0.020)	1.1
	Encephalocele	-0.012 (-0.047, 0.013)	-0.013 (-0.047, 0.012)	1.0
	Spina bifida	-0.002 (-0.018, 0.013)	-0.003 (-0.018, 0.013)	0.8
	Mean of trends for NTD subgroups	-0.005 (-0.105, 0.089)	-0.009 (-0.12, 0.100)	0.5
	SD of trends for NTD subgroups	0.054 (0.001, 0.374)	0.078 (0.002, 0.469)	0.7
6B	Anencephaly	0.002 (-0.014, 0.018)	0.002 (-0.014, 0.019)	1.1
	Encephalocele	-0.012 (-0.046, 0.013)	-0.012 (-0.046, 0.014)	1.0
	Spina bifida	-0.001 (-0.016, 0.013)	-0.001 (-0.016, 0.014)	0.8
	Mean of trends for NTD subgroups	-0.004 (-0.102, 0.093)	-0.006 (-0.112, 0.096)	0.7
	SD of trends for NTD subgroups	0.059 (0.001, 0.391)	0.066 (0.002, 0.425)	0.9

Computational time in Stan was generally longer than in JAGS, with run times of almost 3 times slower for model 5 and up to 37 times slower for model 6A (Appendix Table A13). Effective sample sizes were considerably smaller in Stan, with an average effective sample size of less than 400 compared to over 40,000 in the equivalent JAGS model (Appendix Table A13), indicating even higher levels of autocorrelation in estimates obtained using Stan. A full summary of the posterior distributions for Bayesian models using JAGS and Stan is displayed in Appendix Table A14.

Sensitivity analysis III: use of different prior distributions

Different values for parameters of prior distributions were also considered for the BHM pooled over registry (model 5). Estimates and 95% PCIs for all parameters using 6 combinations of different priors for the means and variances are displayed in Figure 3.10. Combination A shows the parameters as used in models previously presented; the other 5 combinations used a Normal prior for estimation of means and a uniform or half-Cauchy prior for estimation of the variances, as described in the key to Figure 3.10. The estimated mean intercepts and trends in each NTD subgroup and for the overdispersion parameter remained similar for all types of prior considered (see Appendix Table A15 for exact values and related effective sample sizes). When introducing a wider (i.e. vaguer) variance on the uniform prior for the estimation of variances, the 95% PCIs for the mean and SD of the random intercepts became wider (B compared to A). The PCIs for the mean and SD of the random slopes, however (i.e. estimated average trend across the subgroups) became slightly narrower. When using a half-Cauchy distribution for the priors for variance parameters (D, E and F), a higher scale parameter lead to lower precision in the estimates of the mean and SD for the average intercepts and slopes across the NTD subgroups, marked by an increasingly long right tail as the scale parameter increased.



KEY	Prior distribution for means	Prior distribution for variances
F	Normal(0, 0.001)	Half-Cauchy, scale=25
E	Normal(0, 0.001)	Half-Cauchy, scale=10
D	Normal(0, 0.001)	Half-Cauchy, scale=5
C	Normal(0, 0.0001)	Uniform(0, 10)
B	Normal(0, 0.001)	Uniform(0, 100)
A	Normal(0, 0.001)	Uniform(0, 10)

Figure 3.10. Estimated intercepts and slopes in model 5 for neural tube defects with different parameters for prior distributions of means and variances.

3.6.4. Chromosomal anomalies

The estimated total prevalence and 95% CIs for chromosomal CA subgroups within each registry for the latest ten years of data are displayed in Figure 3.11. As seen with the NTDs, the total ten-year prevalence for chromosomal subgroups varied considerably between registries. This was particularly the case for Down syndrome, for which the prevalence ranged from 13.3 cases per 10,000 births in Antwerp up to 31.9 in Paris.

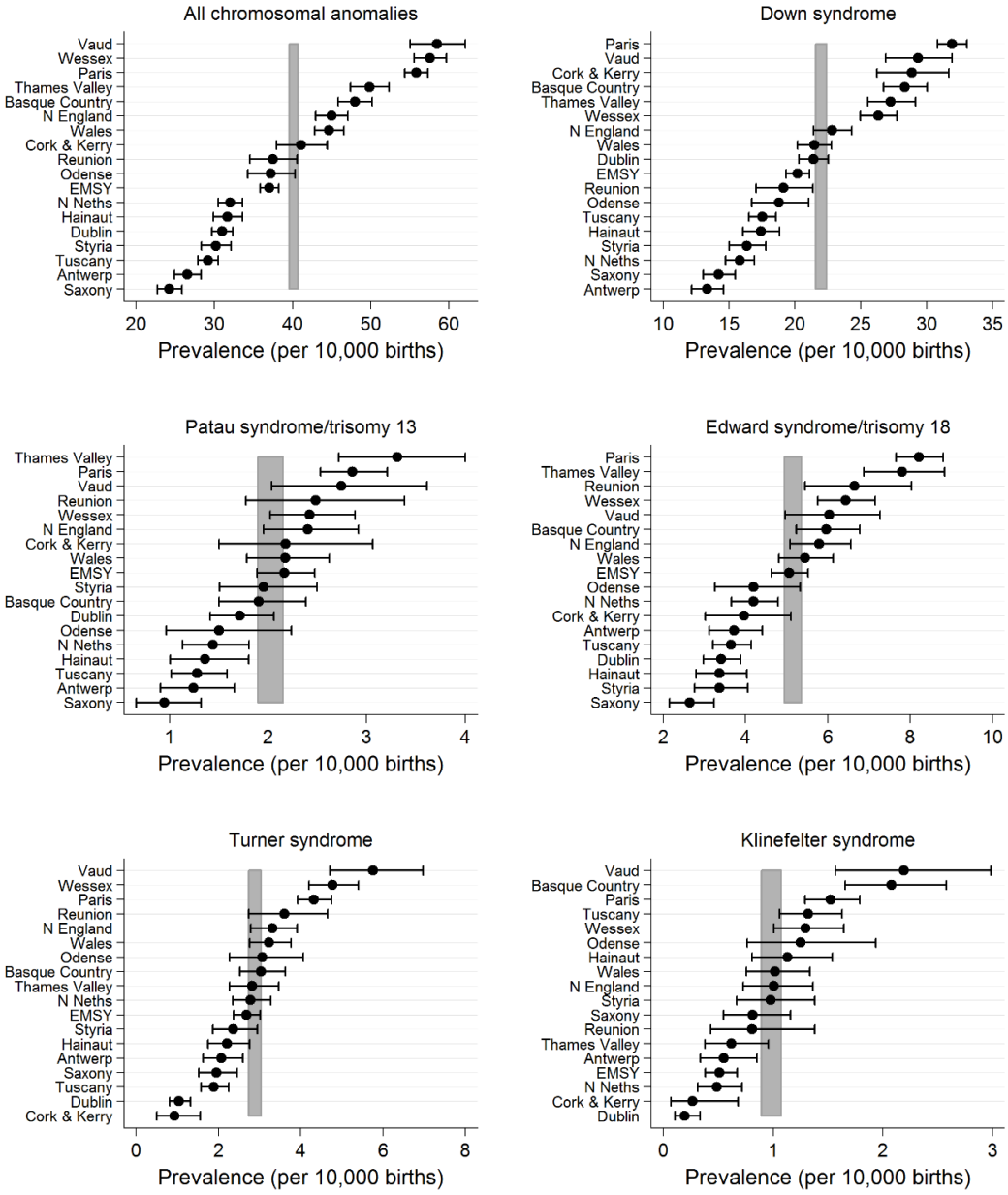


Figure 3.11. Total prevalence of chromosomal anomalies and 95% confidence intervals in 18 EUROCAT registries from 2003 to 2012 from 2003 to 2012, with 99% confidence range for the average prevalence across all registries marked as grey shaded bands for each anomaly.

Figure 3.12 shows the average yearly prevalence and 95% CIs across 18 EUROCAT registries from 2003-2012 in the five chromosomal subgroups, on a log scale such that the difference between two lines on the y-axis represents a doubling in prevalence. Down syndrome was the most common chromosomal CA, with average yearly European-wide prevalence estimated at between 20 and 27 per 10,000 births throughout this time. The other chromosomal subgroups all had considerably lower average yearly prevalence, varying from around 1 per 10,000 births for Klinefelter syndrome to 7 per 10,000 for Edward syndrome. It can be seen in Figure 3.12 that the yearly prevalence of Down syndrome across Europe has increased slightly over this period and that of Klinefelter syndrome has decreased; however, it is not clear whether there has been a similar trend for the other subgroups.

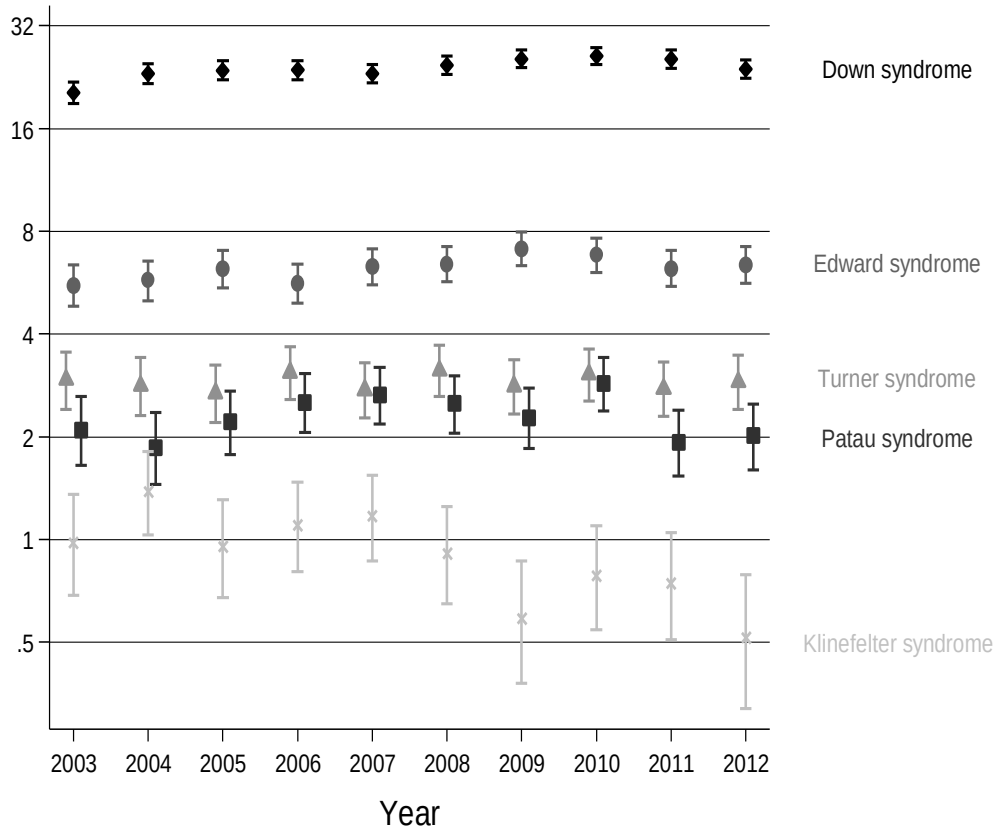


Figure 3.12. Average yearly prevalence and 95% confidence intervals across 18 EUROCAT registries from 2003-2012 in the five chromosomal anomalies.

The estimated trends and 95% CIs or PCIs in the 5 chromosomal subgroups according to the individual models and BHMs are presented in Figure 3.13. There was an increasing trend for Down syndrome and decreasing trend for Klinefelter syndrome, with similar estimates in all models. The estimates for Klinefelter syndrome shrank slightly towards the null when using a BHM, going from an estimated 7.7% average yearly decrease in prevalence in model 1 to the 5.6% estimated by model 6B. There were no changes in the prevalence of Patau or

Turner syndrome for any model. When considering each chromosomal subgroup individually, Poisson regression analyses estimated the increasing trend in Edward syndrome that was nonsignificant after Bonferroni correction to adjust for multiple testing. However, this trend was significant when using each of the BHMs, for which the estimated variability around the average yearly change in prevalence was narrower.

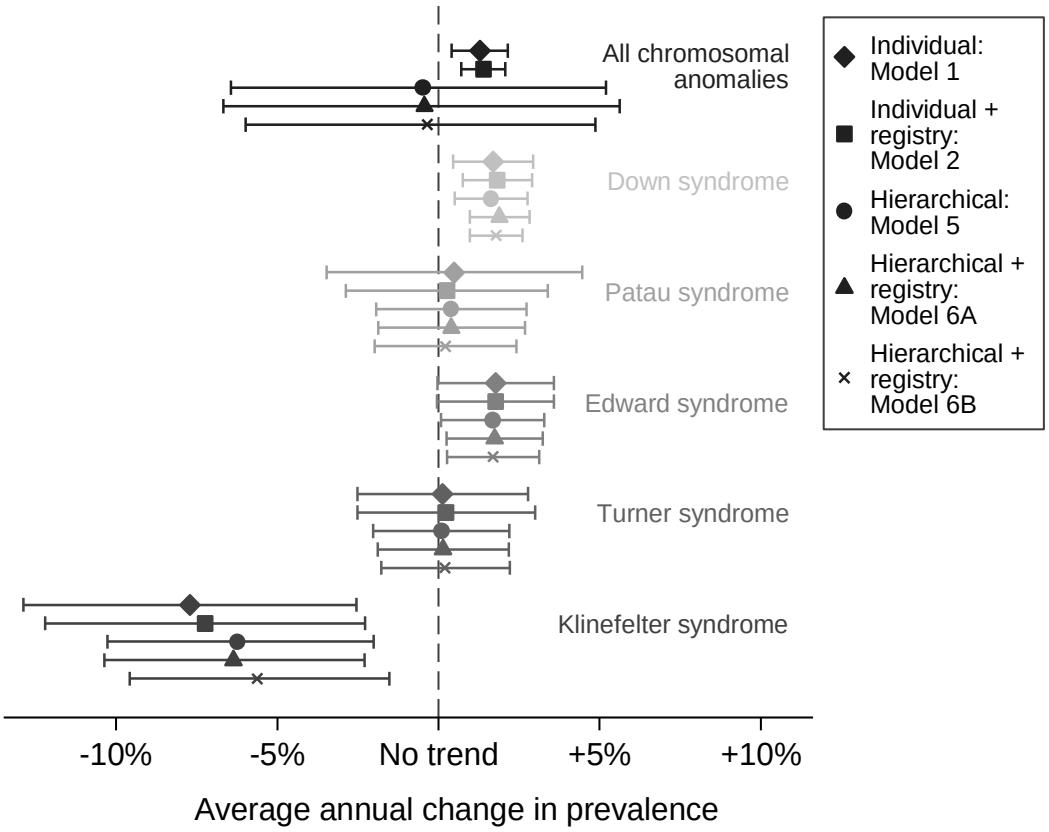


Figure 3.13. Average annual trends in prevalence of chromosomal anomalies; estimates and 95% confidence intervals from individual and hierarchical models as described in section 3.5.3.

Assessment of hierarchical models for chromosomal subgroups

Some key features of model fit for hierarchical models are described in Appendix Table A16. The intercepts and slopes for chromosomal subgroups were highly correlated for models 3 and 4, with the correlation coefficient estimated to be 0.82 and 0.83 respectively. As in the models for NTDs, this could be because of over-parameterisation in the model leading to a very small variance in either random effect, meaning that the model cannot reliably estimate both effects. Some amount of overdispersion was also present for all models, and the estimated standard deviation of the dispersion parameter was smaller for models that pooled information across registry.

Parameters for model 5 and model 6B showed good convergence, with low levels of autocorrelation and good mixing of chains (data not shown). Effective sample sizes were

high for most parameters in the model, with many having an effective sample size close to 60,000 (i.e. the actual total sample size for all three chains after thinning) with the exception of the overdispersion parameter, which had an effective sample size of 3,527 and 881 respectively for models 5 and 6B. As seen in the BHMs for NTDs, the parameters for the trend in each subgroup were well behaved; however, the random intercept parameters for model 6A showed poor convergence and mixing of chains, high autocorrelation and very low effective sample sizes (data not shown).

Hierarchical models for chromosomal subgroups including only autosomal trisomy

Figure 3.14 shows results from the hierarchical models that excluded Klinefelter and Turner syndromes, to evaluate the effectiveness of a model including only the autosomal trisomy subgroups. Estimated yearly trends for Down and Edward syndrome remained almost identical to those obtained in models that had included all five chromosomal anomalies. For Patau syndrome, the model including only the three trisomy subgroups gave a slightly higher estimated trend from just under 0.5% (in model 1) to around a 1% average annual change in prevalence in the three BHMs and with slightly narrower PCIs. This shows how the estimates for Patau syndrome in BHMs were influenced to some extent by significant increasing trends in the other trisomy subgroups; however, these estimates remained statistically nonsignificant for all models considered. Information regarding model fit for estimates displayed in Figure 3.14 is in Appendix Table A17. Model diagnostics were similar to those from models including all 5 subgroups, although the estimated correlation between the intercepts and slopes was perfect in this version of model 3 and 4 ($r=1$ compared to $r\approx0.8$ when including all 5 subgroups).

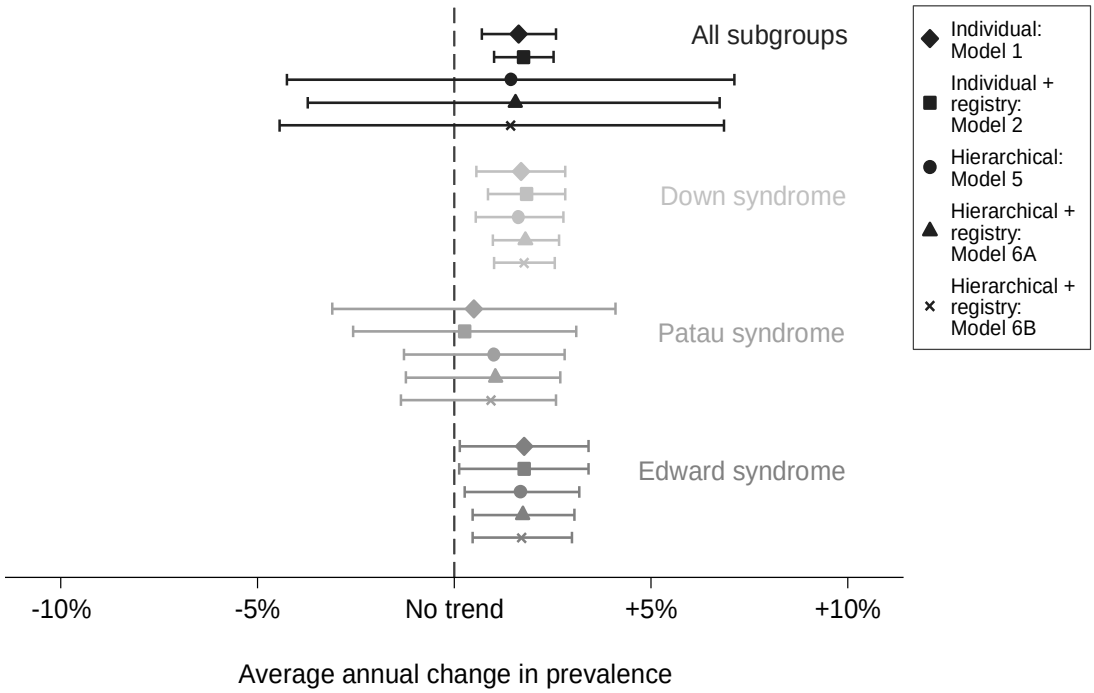


Figure 3.14. Average annual trends in prevalence of autosomal trisomy subgroups.

3.6.5. Digestive system anomalies

Figure 3.15 shows the estimated total prevalence and 95% CIs for the 8 digestive system anomalies for 2003-2012. The total ten-year prevalence was generally more consistent across the 18 registries for the digestive system anomalies compared to NTDs or the chromosomal subgroups, although for come digestive CAs there were a number of registries with estimated prevalence outside the 99% confidence intervals for the European average.

The average yearly European-wide prevalence for each of the digestive anomaly subgroups is displayed in Figure 3.16, ranging from around 0.1 cases per 10,000 live births for annular pancreas to between 2.5 and 3.5 cases per 10,000 for anorectal atresia and stenosis. There was no clear trend for any of the digestive system subgroups from 2003 to 2012, with the possible exception of duodenal atresia or stenosis, which had an estimated prevalence of just over 0.5 per 10,000 births in 2003 rising to 1 per 10,000 births in 2012.

Figure 3.17 shows estimates of the average yearly change in prevalence in digestive system CAs from individual models and BHMs. There was a mix of increasing and decreasing trends across the 8 digestive system subgroups, none of which was significant for any model. When combining the eight anomalies together in a hierarchical model, estimated trends in prevalence were shrank towards the average of the 8 subgroups, which was the line of no

trend as this average is influenced by both positive and negative trends. For subgroups that were estimated with more uncertainty in the individual models (i.e. those with the widest confidence intervals for models 1 and 2), this shrinkage effect was visibly larger.

Assessment of hierarchical models for digestive system subgroups

Information regarding model fit for hierarchical models for digestive system subgroups are described in the Appendix Table A18. The CA intercepts and slopes were almost perfectly correlated for models 3 (estimated correlation coefficient r=0.99) and 4 (r=0.97), again implying that the model is not reliably able to estimate both effects due to over-parameterisation. Some amount of overdispersion was present for all models, with smaller estimated standard deviation of the dispersion parameter for models that did not pool information across registry. The majority of parameters for model 5 and model 6B showed good convergence, low levels of autocorrelation, good mixing of chains and reasonable effective sample size. However, in model 6B the overdispersion parameter had poor mixing of chains and convergence and a very low effective sample size of only 150. As seen in the BHMs for previous groups of anomalies analysed, the random intercept parameters for model 6A showed very poor convergence and mixing of chains, high autocorrelation and low effective sample sizes (data not shown).

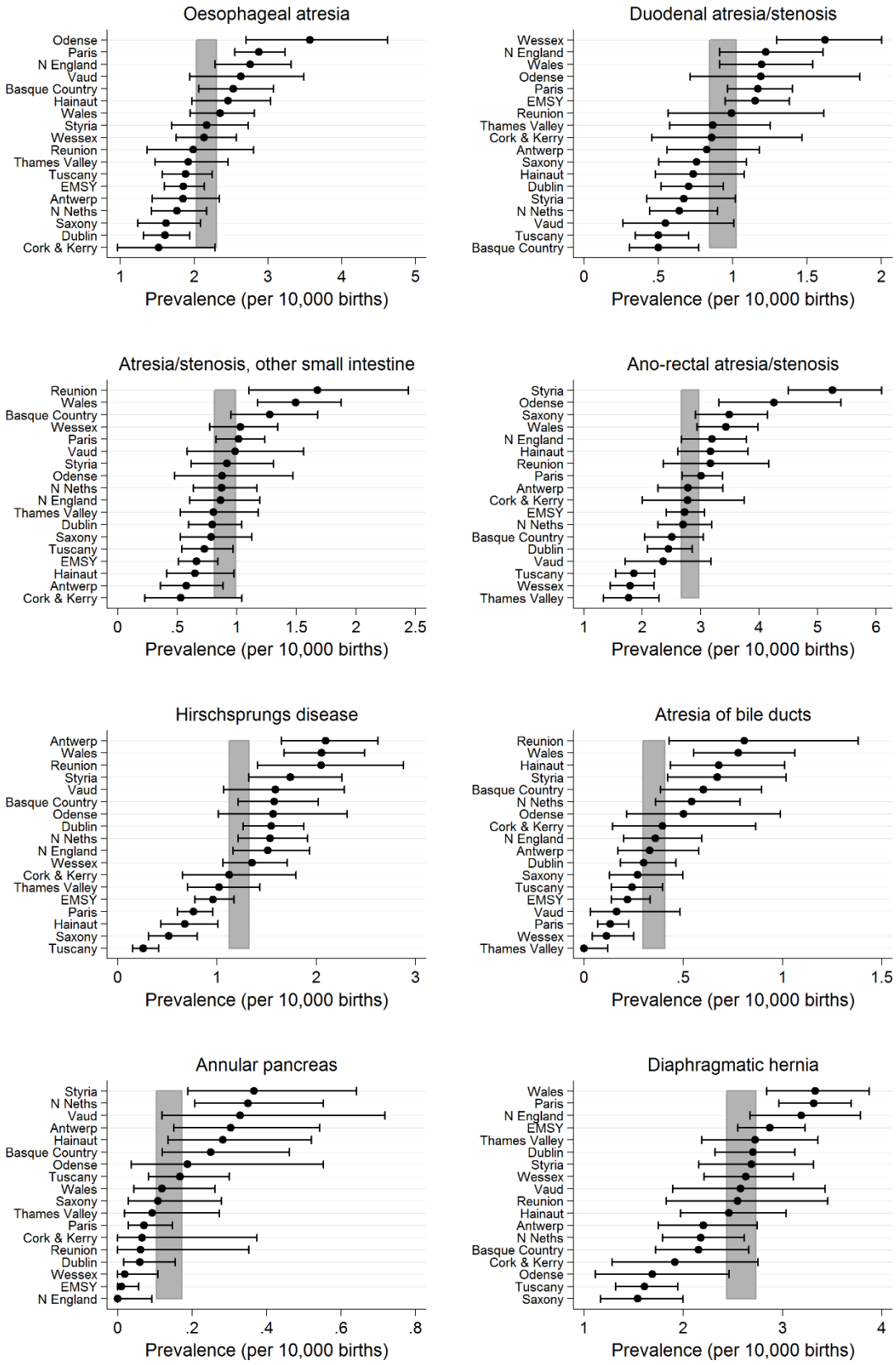


Figure 3.15. Total prevalence and 95% confidence intervals of 8 digestive system CAs in 18 EUROCAT registries from 2003 to 2012, with 99% confidence range for the average prevalence across all registries marked as grey shaded bands for each anomaly.

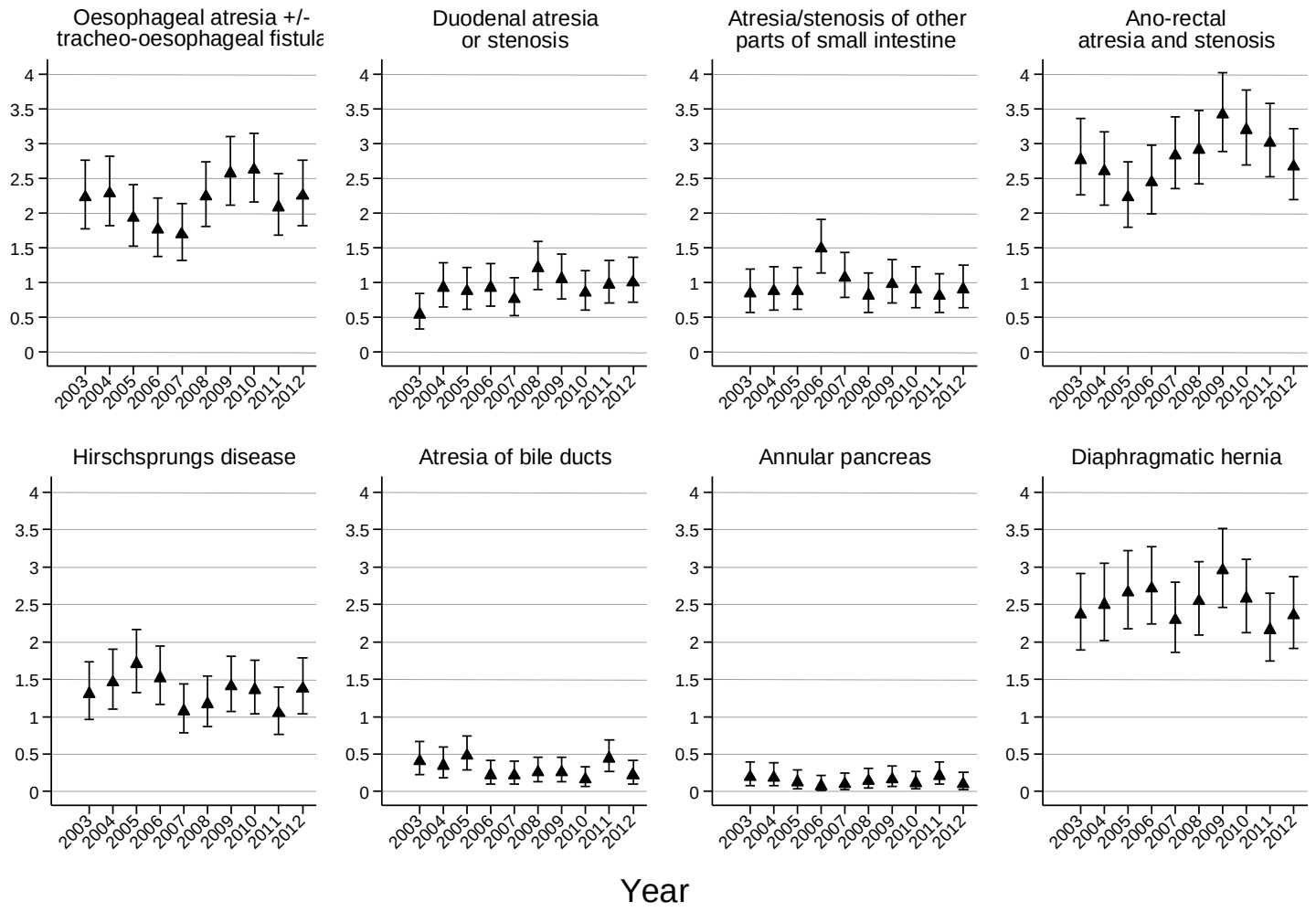


Figure 3.16. Average yearly prevalence of 8 digestive system CAs across 18 EUROCAT registries from 2003-2012.

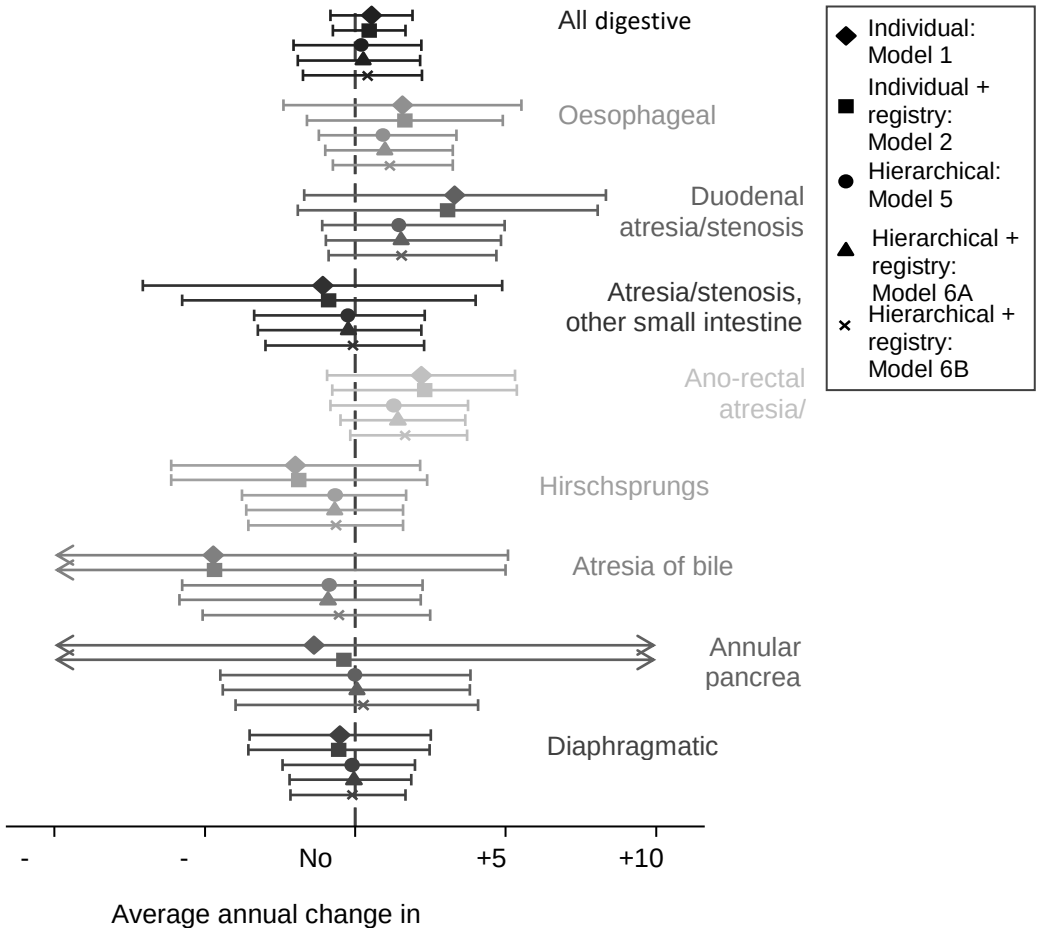


Figure 3.17. Average annual trends in prevalence of digestive system subgroups; estimates and 95% confidence intervals from individual and hierarchical models as described in section 3.5.3.

3.6.6. Congenital heart defects

The estimated prevalence and 95% CIs of the 16 CHDs for 2003-2012 are displayed in Appendix Figure A22 (severity group 1), Figure A23 (severity group 2) and Figure A24 (severity group 3). The total ten-year prevalence was reasonably consistent across the 18 registries for most CHDs in severity groups 1 ("very severe") and 2 ("severe") but varied widely between registries for the more common 3 subgroups in severity group 3. For example, ventricular septal defect had total lowest prevalence in the Wessex registry, with only 9 per 10,000 births recorded, while at the other end of the spectrum was Vaud with almost 55 cases per 10,000 births. Figure 3.18 displays the average prevalence each year across the 18 registries combined for each of the CHD subgroups. Visually it appears that the prevalence might have come down slightly for coarctation of aorta in severity group 2 and atrial septal defect in severity group 3. There also appears to be a slight increase from 2003 to 2012 in the prevalence of tetralogy of Fallot in severity group 2. For the rest of the CHD subgroups there were no clear trends in prevalence according to Figure 3.18.

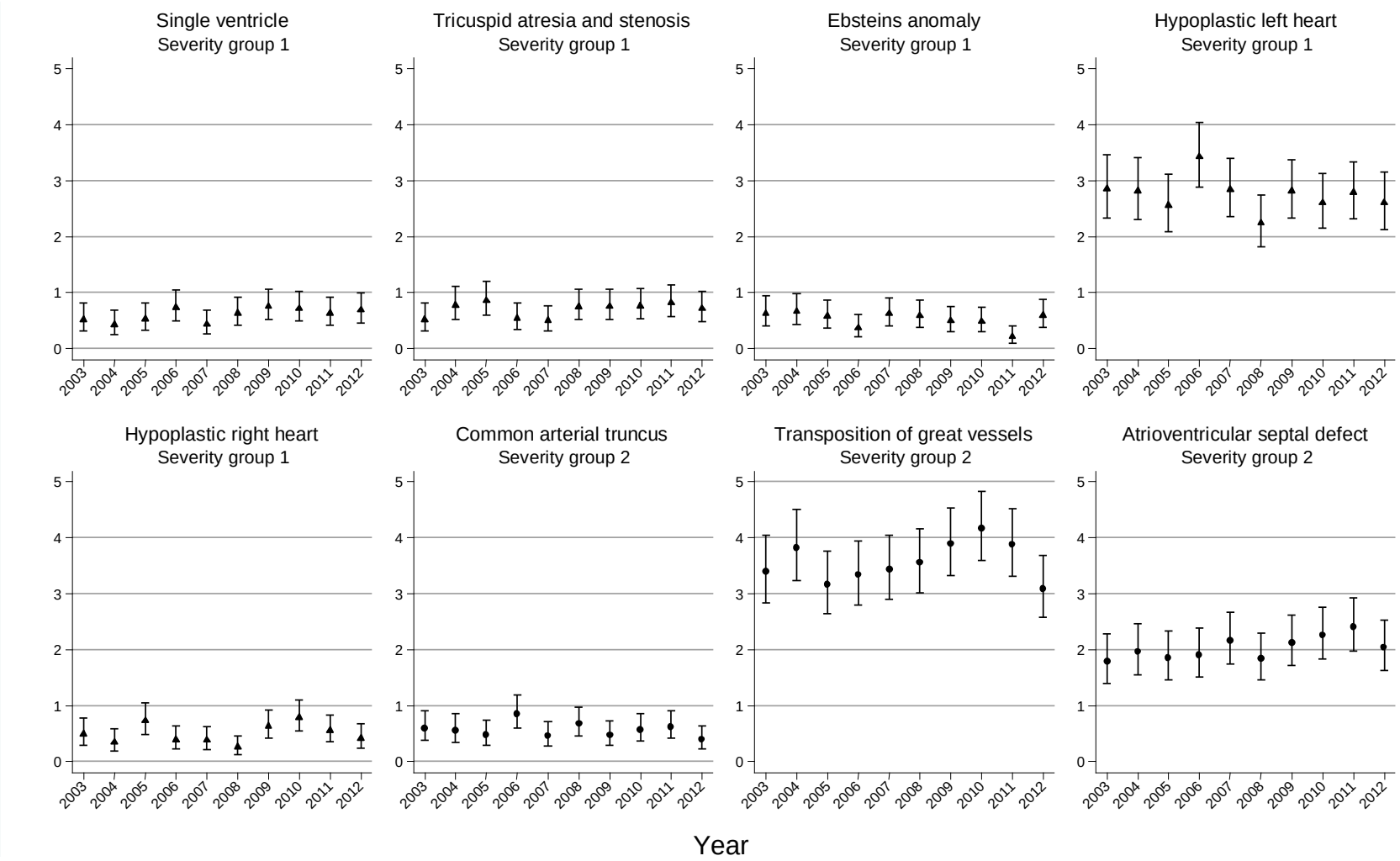


Figure 3.18. Average yearly prevalence across 18 EUROCAT registries from 2003-2012 in the CHD subgroups.

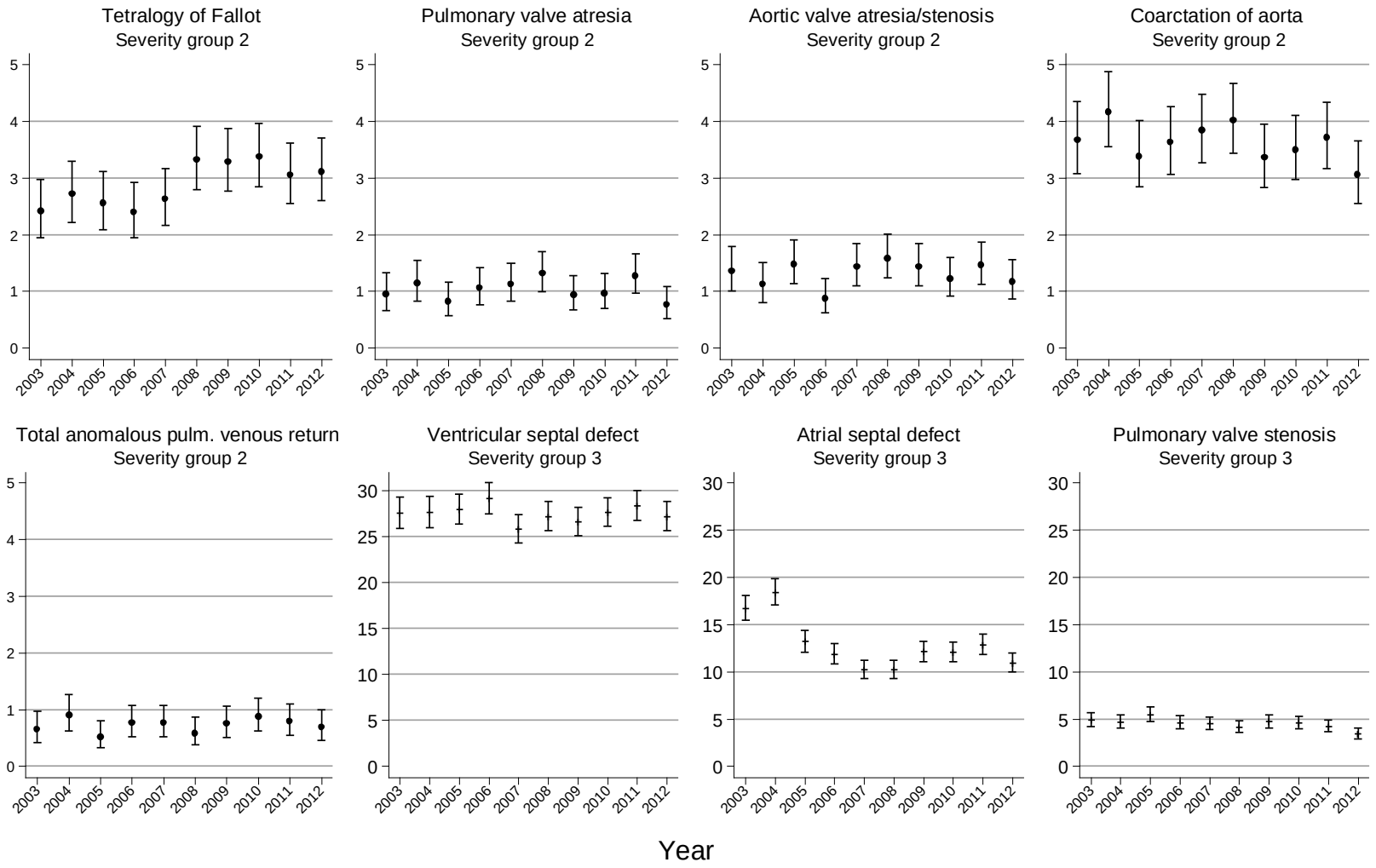


Figure 3.18 (continued). Average yearly prevalence across 18 EUROCAT registries from 2003-2012 in the CHD subgroups.

There was a mix of increasing and decreasing trends in the prevalence of CHD subgroups whether using individual Poisson models (Table 3.6) or BHMs (Table 3.7).

Table 3.6 shows that, after Bonferroni correction to adjust for multiple tests across the 16 CHDs, the only statistically significant estimates from individual models were an increasing trend for tetralogy of Fallot and a decreasing trend for pulmonary valve stenosis, both of which attenuated when including a registry effect in model 2. Three subgroups highlighted in bold showed increasing trends in prevalence in the latest EUROCAT statistical monitoring report; apart from tetralogy of Fallot in model 1, these trends were increasing but nonsignificant in individual models here.

Table 3.7 shows that none of the estimated trends for these 3 CHDs was statistically significant in BHMs pooling the CHDs together. The estimated trends in prevalence shrink towards the average trend across the CHDs in BHMs, which was close to zero due to the influence of both positive and negative trends. Table 3.7Table 3.6 includes estimates from a BHM with an additional random effect for severity group (model 5sev). The decreasing trend for pulmonary valve stenosis seen in model 1 shrank towards the null in all BHMs except model 5sev. On the other hand, models 5, 5sev and 6B all showed a decreasing trend for atrial septal defect, which was not statistically significant in individual models after Bonferroni adjustment. There were no significant trends observed for any CHD subgroup in model 6A.

Table 3.6. Estimated average annual trends in 16 congenital heart defects from individual models.

Severity group	Congenital heart defect	Model 1: Individual models pooled over registry	Model 2: Individual models with a random effect for registry
1	**Single ventricle [a]**	**0.041 (-0.025, 0.108)**	**0.046 (-0.020, 0.112)**
	Tricuspid atresia and stenosis	0.022 (-0.040, 0.084)	0.020 (-0.043, 0.083)
	Ebstein's anomaly	-0.043 (-0.119, 0.033)	-0.042 (-0.113, 0.029)
	Hypoplastic left heart	-0.010 (-0.044, 0.023)	-0.010 (-0.043, 0.022)
	Hypoplastic right heart	0.021 (-0.086, 0.128)	0.020 (-0.057, 0.097)
2	Common arterial truncus	-0.019 (-0.091, 0.052)	-0.021 (-0.095, 0.053)
	Transposition of great vessels	0.008 (-0.022, 0.038)	0.007 (-0.022, 0.037)
	Atrioventricular septal defect [a]	**0.023 (-0.013, 0.059)**	**0.022 (-0.019, 0.063)**
	Tetralogy of Fallot [a]	**0.034 (0.003, 0.064)[b]**	**0.032 (-0.001, 0.064)**
	Pulmonary valve atresia	-0.001 (-0.056, 0.054)	0.002 (-0.056, 0.059)
	Aortic valve atresia/stenosis	0.006 (-0.044, 0.057)	0.012 (-0.039, 0.062)
	Coarctation of aorta	-0.014 (-0.041, 0.013)	-0.015 (-0.048, 0.017)
	Total anomalous pulmonary venous return	0.010 (-0.051, 0.070)	0.010 (-0.050, 0.071)
3	Ventricular septal defect	-0.001 (-0.012, 0.009)	0.003 (-0.014, 0.020)
	Atrial septal defect	-0.041 (-0.087, 0.006)	-0.025 (-0.058, 0.008)
	Pulmonary valve stenosis	-0.028 (-0.054, -0.001)[b]	-0.019 (-0.049, 0.010)
-	All subgroups	-0.010 (-0.017, -0.002)	-0.005 (-0.016, 0.006)

[a] Rows in bold indicate subgroups that showed significant increasing trends in prevalence in the 2012 EUROCAT statistical monitoring report (published 2015)

[b] "Significant" trends after Bonferroni correction

Table 3.7. Estimated average annual trends in 16 congenital heart defects from Bayesian hierarchical models.

Severity group	Congenital heart defect	Hierarchical model: Pooled over registry (model 5)	Hierarchical model: Pooled over registry, with severity grouping (model 5sev)	Hierarchical model: Random effect for registry (model 6A)	Hierarchical model: Random effect for registry (model 6B)
1	**Single ventricle** [a]	**0.016 (-0.015, 0.053)**	**0.023 (-0.013, 0.069)**	**0.009 (-0.013, 0.043)**	**0.016 (-0.011, 0.050)**
	Tricuspid atresia and stenosis	0.009 (-0.021, 0.042)	0.013 (-0.020, 0.053)	0.005 (-0.017, 0.033)	0.008 (-0.019, 0.038)
	Ebstein's anomaly	-0.017 (-0.054, 0.015)	-0.019 (-0.070, 0.019)	-0.006 (-0.037, 0.016)	-0.012 (-0.045, 0.015)
	Hypoplastic left heart	-0.007 (-0.030, 0.015)	-0.006 (-0.032, 0.018)	-0.003 (-0.025, 0.016)	-0.006 (-0.027, 0.013)
	Hypoplastic right heart	0.007 (-0.025, 0.042)	0.012 (-0.025, 0.056)	0.004 (-0.019, 0.033)	0.006 (-0.022, 0.038)
2	Common arterial truncus	-0.009 (-0.042, 0.023)	$1.2x10^{-8}$ (-0.034, 0.025)	-0.003 (-0.032, 0.019)	-0.006 (-0.036, 0.022)
	Transposition of great vessels	0.005 (-0.017, 0.027)	0.007 (-0.012, 0.027)	0.002 (-0.016, 0.024)	0.004 (-0.015, 0.023)
	Atrioventricular septal defect [a]	**0.014 (-0.011, 0.040)**	**0.013 (-0.007, 0.038)**	**0.007 (-0.012, 0.033)**	**0.012 (-0.009, 0.036)**
	Tetralogy of Fallot [a]	**0.022 (-0.003, 0.048)**	**0.019 (-0.002, 0.045)**	**0.009 (-0.009, 0.037)**	**0.018 (-0.003, 0.041)**
	Pulmonary valve atresia	-0.001 (-0.029, 0.027)	0.004 (-0.022, 0.027)	0.0003 (-0.023, 0.024)	0.001 (-0.024, 0.027)
	Aortic valve atresia/stenosis	0.003 (-0.023, 0.030)	0.006 (-0.017, 0.029)	0.004 (-0.017, 0.029)	0.005 (-0.018, 0.031)
	Coarctation of aorta	-0.010 (-0.032, 0.011)	-0.003 (-0.027, 0.016)	-0.005 (-0.028, 0.013)	-0.010 (-0.029, 0.009)
	Total anomalous pulmonary venous return	0.003 (-0.027, 0.035)	0.007 (-0.019, 0.034)	0.003 (-0.020, 0.029)	0.004 (-0.022, 0.033)
3	Ventricular septal defect	-0.001 (-0.018, 0.016)	-0.004 (-0.024, 0.016)	0.001 (-0.015, 0.019)	0.002 (-0.010, 0.015)
	Atrial septal defect	-0.032 (-0.052, -0.011) [b]	-0.038 (-0.058, -0.017) [b]	-0.009 (-0.031, 0.008)	-0.022 (-0.038, -0.005) [b]
	Pulmonary valve stenosis	-0.020 (-0.043, 0.001)	-0.026 (-0.049, -0.004) [b]	-0.007 (-0.031, 0.011)	-0.012 (-0.032, 0.006)
	All subgroups	-0.001 (-0.015, 0.013)	-0.001 (-0.150, 0.147)	0.001 (-0.010, 0.013)	0.0005 (-0.011, 0.013)

[a] *Rows in bold indicate subgroups that showed significant increasing trends in prevalence in the 2012 EUROCAT statistical monitoring report (published 2015)*

[b] *"Significant" trends (95% PCI does not include 0)*

Adding a random effect for severity group in hierarchical models for congenital heart defects

When adding random effects for the severity subgroup indicator to the hierarchical model for CHDs, decreasing average yearly trends were observed for two of the CAs in severity group 3 (atrial septal defect and pulmonary valve stenosis; Figure 3.19, Table 3.6). There were no significant trends for any other CHDs or in the average prevalence across each severity group. Estimates for each severity group shrank slightly towards the average of that group rather than just towards the overall (null) average across all 16 subgroups. All estimates from model 5sev are shown in Appendix Table A19.

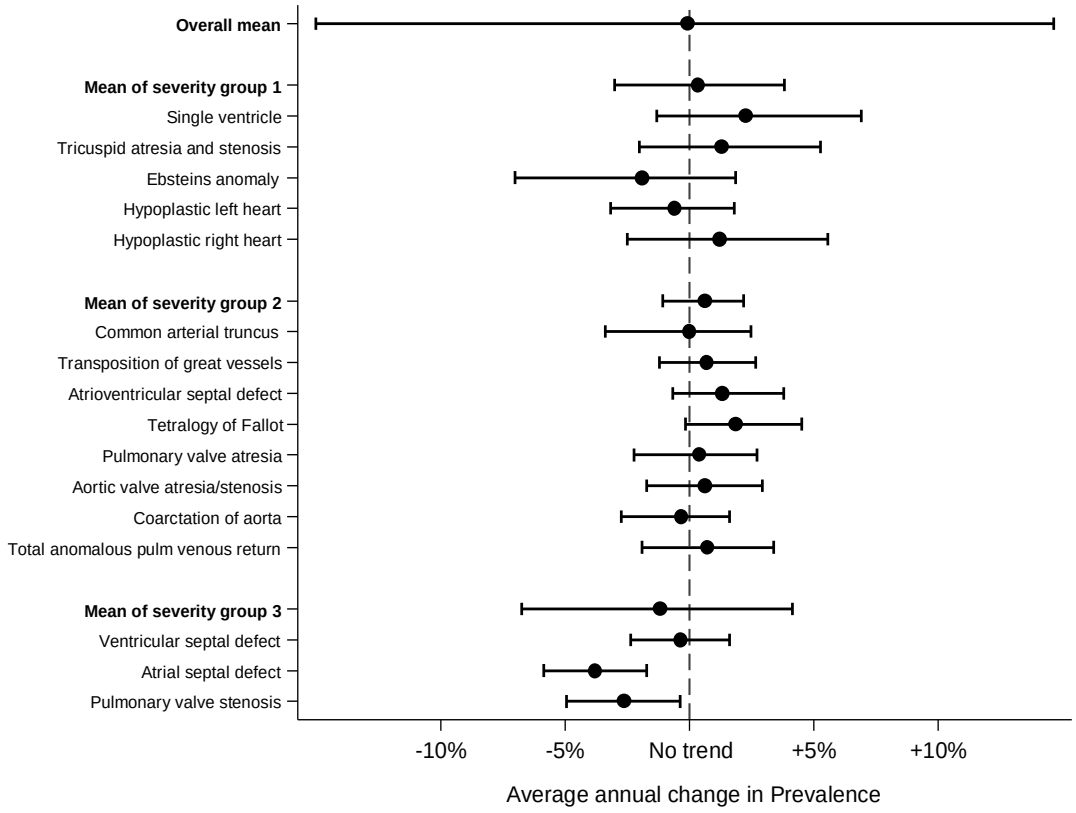


Figure 3.19. Estimated average annual trends in 16 congenital heart defects in a hierarchical model with an additional term for severity subgroup.

Assessment of hierarchical models for congenital heart defects

Information regarding model fit in the hierarchical models for CHD subgroups is presented in the Appendix Table A20. The intercepts and slopes were negatively correlated for CAs in models 3 (estimated correlation coefficient r=-0.47) and 4 (r=-0.65), as in previous models indicating that the model cannot estimate both effects, likely due to very small variance in the estimated trends caused by over-parameterisation. Some amount of overdispersion was present for all models, with smaller estimated standard deviation of the dispersion parameter for models that did not pool information across registry. All parameters for

model 5 and model 6B showed good convergence, low levels of autocorrelation, good mixing of chains and generally high effective sample size. For model 5 when including an additional term for severity group, all parameters showed good convergence and effective sample size, although the group level parameters (i.e. the estimated intercepts and slopes for each severity group, and for the group of CHDs overall) in this model showed less consistent mixing of chains (data not shown). There was high autocorrelation in the estimated SD parameter for each of the severity group trends. Once again, the random intercept parameters for model 6A showed very poor convergence and mixing of chains, very high autocorrelation and low effective sample sizes.

3.7. Discussion

In this chapter, BHMs were applied to EUROCAT data to assess the effect of modelling group effects for CAs and/or registries in analyses of European trends in the prevalence of CAs. Standard EUROCAT coding was used to group CAs together. BHMs that grouped together similar CAs were compared to models considering each CA separately, and results were found to be broadly similar for the four different groups of CAs considered.

3.7.1. The Poisson regression model

Poisson regression was used to model yearly counts of cases in each CA throughout all analyses in this chapter; therefore, the assumptions of Poisson regression needed to be considered, in particular that of equality of the mean and variance. Using the NTD encephalocele as an example, various models were considered to investigate potential overdispersion in the data, and to assess the effect of this on estimated changes in prevalence. However, throughout the different models in this chapter there was no evidence of overdispersion when considering trends over the past ten years of data. It is worth noting that since only ten years of data have been considered (to estimate recent trends) there may be insufficient data points to provide a reliable estimate of the amount of overdispersion in a model. However, estimates of the average yearly changes in prevalence of encephalocele and of the dispersion parameter were very similar for all models considered. The Poisson-lognormal model was therefore used for all other analyses in this chapter, since this model allowed for potential overdispersion in the data whilst being equivalent to the standard Poisson model when no overdispersion was present. In practice, the Poisson-lognormal model was also straightforward to implement for hierarchical models, in particular compared to approaches such as quasi-Poisson or negative binomial models.

3.7.2. Ten year trends in congenital anomalies analysed for this chapter in the context of previously published studies

Estimated trends in prevalence were similar for all types of CA considered in this chapter, whether considering CAs separately in individual models or together in hierarchical models. Identified trends were largely as expected and consistent with those observed in other studies. Increasing trends in chromosomal anomalies were observed, which are known to be due to maternal age and changes in prenatal screening practices [Cocchi et al., 2010, Loane et al., 2013]. The decreasing trend in Klinefelter syndrome was also observed in the three most recent EUROCAT statistical monitoring reports. It was suggested by the authors that this decrease is more likely to reflect changes in prenatal screening than a true decline in prevalence, as there have been less invasive prenatal tests as a result of the first trimester screening test being implemented [EUROCAT Central Registry, 2014a]. NTD prevalence remained stable in EUROCAT registries, as has been observed elsewhere [Botto et al., 2006a, Khoshnood et al., 2015]. This might be explained by the lack of folic acid fortification in Europe and poor uptake of folic acid supplementation; in the UK, for example, less than 30% of women took folic acid prior to their pregnancy in 2011–2012 [Bestwick et al., 2014]. Prevalence in three of the digestive system subgroups was found to be significantly increasing in the latest EUROCAT statistical monitoring report [EUROCAT Central Registry, 2015]. A similar estimated increase in prevalence in these three subgroups was observed here, although these trends did not reach statistical significance in independent models. A smaller number of EUROCAT registries were included in the dataset for this thesis, hence the lack of statistical significance in these analyses may be due to a relative lack of power. Increases in the prevalence of the CHD tetralogy of Fallot (severity group 2) was consistent with previous findings [EUROCAT Central Registry, 2015]. Although not reaching statistical significance, estimated increasing trends for single ventricle (severity group 1) and atrioventricular septal defect (severity group 2) were also consistent with the EUROCAT findings in terms of the direction of effect. Estimated decreases in prevalence of atrial septal defect and pulmonary valve stenosis (severity group 3), however, were not consistent with those observed in other studies, where either no significant changes or increasing trends have been previously observed [van der Linde et al., 2011, EUROCAT Central Registry, 2015]. For pulmonary valve stenosis, the decreasing trend was only observed in individual models pooling over registry (model 1) and a BHM additionally including a grouped effect for severity (model 5sev), but this trend was not significant in other models (see Table 3.6 and Table 3.7). Published prevalence estimates for CHDs are

also known to vary substantially due to differing definitions of cases across studies. It is likely that the differences in estimated trends here reflect changes in reporting for these CAs rather than real changes in prevalence; since August 2007, for example, EUROCAT coding has required that only atrial septal defect cases that have been confirmed as still present after 6 months of age be reported [EUROCAT Central Registry, 2013]. Differing prenatal screening practices in the particular set of registries with data available for this thesis may also have influenced the observed trends [Hoffman and Kaplan, 2002, Garne et al., 2012b, Baardman et al., 2014].

3.7.3. Performance of Bayesian hierarchical models

Many of the BHMs applied to groups of CAs in this chapter were over-parameterised, in particular when taking the registry effect into account, as this greatly increased the number of parameters being estimated by each model. Gibbs sampling is known to be inefficient for models in which the parameters are highly correlated, as this was seen here for the random effects parameters in hierarchical models. This was especially noticeable for smaller groups of CAs that only had a small number of random effects levels, e.g. hierarchical models grouping the 3 NTDs. The use of Stan was therefore investigated to see if its alternative samplers could provide a solution for parameters where the JAGS samplers did not converge. However, results from models using Stan were found to be very similar to those obtained using JAGS, and estimated trends in each CA remained similar across the different models whether using JAGS or Stan. For example, for all groups of CAs considered in this chapter, the random intercept parameters for model 6A showed poor convergence and mixing of chains, very high autocorrelation and low effective sample sizes. The over redundant information present in these models is illustrated in the trace plots for these model parameters; for example, in Figure A8 the trace plot for each chain (e.g. indicated by a different colour line) for the estimated random intercepts for each NTD balances with the corresponding chain for the registry intercepts in this model, shown in Figure A9 (i.e. the green line in Figure A8 is always low for each NTD, whilst the green line in figure A9 is always high). It is likely that this model does not converge because the second intercept parameter u_j is balancing with the first u_{0k} (see Table A4 for notation), whereas in model 6B there is only one parameter u_{0jk}, which achieves better convergence. Table A14 shows how both JAGS and Stan balance the estimates for the registry and the CA intercepts in model 6A (in JAGS the u_j take larger negative values than the u_{0k}, and it is the opposite way around in the same Stan model).

The use of different parameters for prior distributions were also considered for hierarchical models for NTDs. As expected, the estimated group means of random effects in these models had wider PCIs for larger values of the prior variance. The upper values for the 95% PCI of the estimated standard deviation for random effects parameters may start to be considered unrealistically high when the prior variance is allowed to take very large values. Estimated trends in prevalence of each NTD subgroup remained stable, however, across all choices of parameters considered. The amount of shrinkage of these estimates did not appear to be affected by the choice of prior parameter values. In addition, the estimated overdispersion parameter remained stable across the six different prior formulations considered here (see Figure 3.10).

Varying levels of correlations were observed between the random effects, with perfect (negative or positive) correlation being estimated by frequentist hierarchical models for NTDs and the chromosomal trisomy subgroups, indicating that there was redundant information in these models. With only 3 subgroups for NTDs and the chromosomal trisomies (and therefore 3 levels for the random effect of CA), these very small groups of related CAs likely had too few levels to reliably estimate the random effects parameters in BHMs. Indeed, it is well known that a random effects model with less than five levels for the random effect parameter does not perform well, with such models often showing poor convergence and over-parameterisation [Greenland, 2000, Gelman Andrew and Hill, 2007]. Some of the hierarchical models in this chapter were effectively more complex than the data could support in practice, especially when including the effect of registry in the model. Hierarchical models that did not include the effect of registry performed better in terms of diagnostic model checks. It might therefore be inadvisable to include the effects of registry at the first stage of analysis, but rather identify CAs that show potential changes in prevalence and then further adjust and/or stratify for registry effects only for those CAs where a potential trend has been highlighted. However, this approach could possibly lead to important changes in prevalence that are country or region specific being missed, for example if due to some environmental hazard in one particular area that would not likely be present in the data overall when pooled across many registries.

3.7.4. Use of hierarchical models in the analysis of congenital anomaly data

Hierarchical models have proven useful in the field of pharmacovigilance, where they have been used in the detection of potential adverse drug reactions [Berry and Berry, 2004, Xia et al., 2011, Crooks et al., 2012]. Natural hierarchies in drug and adverse event coding have

been used to group similar drugs or AEs together, such that models for each drug-adverse event combination incorporate information from analyses of other similar drugs and AEs [Prieto-Merino et al., 2011]. In this chapter, the same rationale was applied to CAs; however, the situation differs to that for adverse drug reactions, where the hierarchical classification systems may provide more natural hierarchies than the grouping of anomalies according to the defined subgroups. In practice, analyses in this chapter have shown that the EUROCAT subgroup coding hierarchy provides groups of CAs that, whilst similar in certain aspects, are still too heterogeneous to be grouped together sensibly when analysing changes in prevalence. This is because the shrinkage in BHMs will largely pull estimates towards the null if there is a mixture of increasing and decreasing trends, as for CHD and digestive CAs, for example. It is therefore possible that potential changes in prevalence in analyses of heterogeneous groups of CAs such as these could actually be masked by hierarchical models.

On the other hand, this shrinkage can help control estimates based on small counts by including information from the rest of the group. Moreover, this can be thought of as a natural "penalisation" if we consider that a hierarchical model is simultaneously looking for changes in prevalence for a number of subgroups, compared to individual models where this multiple testing aspect is not taken into account (and a number of false positive results are therefore likely). Indeed, the use of hierarchical models has been recommended as a natural way of accounting for multiple comparisons. Gelman Andrew et al. [2012] advised that hierarchical models give better results in general than basic multiple comparisons corrections, and furthermore are not more difficult to implement than some of the more complex classical multiple comparisons correction procedures. For a group where the mean trend across subgroups is close to the null, this penalisation will mean that the estimated trend is no longer a "signal" in the hierarchical model, for example as seen for the CHD tetralogy of Fallot in severity group 2 (Table 3.6 and Table 3.7Table 3.6). For a group where the mean trend is not so close to the null, however, this penalisation might actually lead to an increase in the strength and/or precision of a signal, for example for atrial septal defect in severity group 3 (Table 3.6 and Table 3.7). Furthermore, the same signal might be reduced or enhanced depending on which grouping is used; for example, the trend in pulmonary valve stenosis attenuated if considering all CHD groups together, but was maintained when also including the severity grouping in the model (Table 3.6, Table 3.7, Figure 3.19). These examples highlight that the posterior distribution is sensitive to

informative prior information, which here is influenced by the way the groups have been defined (rather than the use of informative parameters for prior distributions).

In this chapter, EUROCAT subgroups that were considered related (being in the same organ system class) were found to vary considerably in terms of their differing proportional yearly changes in prevalence. There might be other CAs not considered in these analyses for which it might potentially be more relevant or useful to analysed together in a hierarchical model. Specific codes within EUROCAT subgroups could potentially be grouped together (such that the EUROCAT subgroup would itself become the group of similar codes), however these would likely contain too few cases to perform meaningful statistical analyses. There are also known relationships between CAs that lie within different groups of the EUROCAT hierarchy, producing a further level of complexity. In addition to NTDs, for example, there are a number of other anomalies across different body systems that are thought to be sensitive to folate levels during pregnancy, including CHDs, clefts and limb reduction defects [Wilson R. D. et al., 2015]. If there were evidence that folate levels had been increasing in Europe, then it might have been useful to analyse all these anomalies together in a hierarchical model. However, from examining the NTDs alone here (and in other studies), no such change has occurred in Europe and hence such models were not considered useful to investigate further. Similarly, EUROCAT now includes a VATER/VACTERL association subgroup that comprises anomalies of the vertebra, anal atresia, CHDs, trachea-oesophageal fistula, oesophageal atresia, radial anomaly and limb defect, which are known to occur together more frequently than expected by chance. However, the heterogeneity of trends observed in just the CHD component of this subgroup indicates that hierarchical models are not likely to add useful information to such an analysis.

Another way of potentially improving the BHMs used in this chapter might be to include better (i.e. more informative) prior information in to the analysis. For example, the parameter for the intercept u_{0k} in model 5 determines the baseline prevalence rate in that model, and it could be possible to include information about this parameter from known prevalence data in a relatively easy way, by centring the prior distribution for the parameter u_{0k} on the known baseline (or average) prevalence of the group of CAs being considered in the model. Use of more informative priors in this way might potentially strengthen the BHMs in these analyses.

When examining changes in CA prevalence there are multiple factors that can have an influence, such as reporting, case ascertainment or screening practices. Hierarchical models

might therefore be more relevant when considering the risks of specific exposures in relation to the prevalence of CAs. One such important risk factor for CAs that is routinely monitored is medication use during early pregnancy, and the use of hierarchical models for analysis of these is explored in the following three chapters of this thesis.

3.7.5. Strengths and limitations of EUROCAT data

EUROCAT registries collect data that is ascertained from multiple sources and includes information on all major structural and chromosomal CAs [Boyd et al., 2011, Loane et al., 2011a], providing high quality population-based data across multiple European countries and allowing the inclusion of a large number of CA cases covering over four million births over ten years for these analyses. EUROCAT registries include information on cases of prenatal diagnosis followed by termination of pregnancy, enabling the inclusion of cases that would otherwise have gone undiagnosed, or unreported amongst spontaneous abortions or stillbirths. EUROCAT have a detailed data quality strategy, which includes the development and annual update and monitoring of a set of data quality indicators relating to both diagnostic and registry processes [Loane et al., 2011a]. Despite this, the possibility that unknown data artefacts might be responsible (or partly responsible) for any observed changes in prevalence (or lack of observed trends) cannot be excluded. However, since the aim of this chapter was to compare statistical methods, the presence of any such effects would likely affect all models considered in a similar way.

A potential limitation of these analyses is that it was not possible to include data from all of the EUROCAT member registries; hence, some trends that were seen in the latest statistical monitoring report did not reach statistical significance here, likely due to the smaller sample sizes included. However, it does not seem probable that increasing the sample size would have improved the performance of hierarchical models, since the issues were more related to CA group sizes and difficulty in forming larger groups that were sufficiently homogeneous such that grouping in a hierarchical model was useful.

3.7.6. Summary and Conclusions

Hierarchical models considered here did demonstrate how sharing information between subgroups of anomalies can provide a sensible “penalisation” to help avoid false positive signals by shrinking estimated trends towards the null when there is no evidence of other trends in the rest of the group, whilst maintaining signals of changes in prevalence when there are others in the group. Hierarchical models using the EUROCAT hierarchy of CA subgroups, however, presented no substantial improvements over the independent

analyses of each subgroup. When using EUROCAT subgroups for analysis, therefore, considering each CA separately remains an appropriate method for the detection of potential changes in prevalence by relevant surveillance systems. Findings from this chapter have formed the basis of a journal article published in “Birth Defects Research Part A: Clinical and Molecular Teratology” [Cavadino et al., 2016], which is presented in Appendix A8.

Chapter 4: Medication use during pregnancy and the associated risk of congenital anomalies: review, methods of model comparison and description of EUROmediCAT data.

4.1. Introduction

Pregnant women are excluded from the majority of safety studies for new medications, so little is known about the potential risks to a foetus for most medications. It is therefore important that statistical methods used in surveillance analyses of first trimester medication use and the related risks of CAs can provide pregnant women with access to the most up to date and relevant information regarding potential safety concerns that may arise. This chapter begins by reviewing current statistical methods for the systematic detection of harmful medications, both in general and more specifically for teratogenic medications during early pregnancy and the associated risk of CAs. Potential improvements to these methods are then discussed, which aim to incorporate information in the analysis about similarities amongst groups of medications or CAs. Difficulties in the evaluation of signal detection methods for CA data are discussed, and the Australian classification system for prescribing medicines in pregnancy is presented as a way of comparing the methods explored in chapters 5 and 6. The EUROmediCAT dataset used in chapters 5 and 6 is then described and summarised.

4.2. Review of methods used to identify potentially harmful medications and rationale for new approaches to the analysis of EUROmediCAT data

This section introduces signal detection (section 4.2.1) and the statistical methods that have been used in the analysis of potentially harmful medications. This is first discussed in the context of large Spontaneous Reporting (SR) databases of suspected adverse drug reactions, an area in which there has been a great amount of research and methodological developments (section 4.2.2). Whilst some of these databases do include cases of CA, this information is generally limited and has not been routinely analysed. Routine signal detection analyses of population-based CA data has only been initiated and developed in recent years, with the setting up of a European network of CA registries that collect data on medication exposures during pregnancy. The statistical methods currently used to perform these routine CA analyses are then discussed (section 4.2.3).

4.2.1. What is signal detection?

A "signal" for medication safety is defined by the World Health Organisation as '*reported information on a possible causal relationship between an adverse event (AE) and a medication, the relationship being unknown or incompletely documented previously*' [Edwards and Biriell, 1994]. Note the use of the word "possible", highlighting that a signal is not evidence of a causal relationship, but rather a warning sign that requires further investigation. With thousands of reports in any medication safety database, it is clearly not feasible to assess each individual report separately; quantitative methods of signal detection are needed in order to focus efforts to enable more detailed medical review on likely true signals. The main aims of quantitative signal detection are summarised by Bate and Evans [2009] as follows

- *to flag potential signals that might be missed*
- *to prioritise resources for signal detection when combined with more traditional methods, focussing clinical review on the most likely candidates*
- *to detect more complex dependencies in the data, which are hard to detect by manual review, in particular drug-interactions*
- *to aid prioritisation of signals* [Bate and Evans, 2009, p.427]

Signal detection is first step in a wider signal management process, which includes the follow up and assessment of signals in detailed literature searches and the collection of additional information, followed by the communication of resulting recommendations for action to all stakeholders involved, such as international drug monitoring organisations or relevant drug companies [European Medicines Agency, 2012].

4.2.2. Statistical methods used in the analysis of spontaneous reporting data

Pharmacoepidemiology is the study of the distribution and determinants of drug-related events, and the application of such studies to promote safe and effective drug treatment practices. Large SR databases have been set up with the aim of detecting signals of AEs by searching for drug-AE combinations that have unexpectedly high numbers of reports. These are determined using quantitative methods based on measures of "disproportionality", which aim to identify drug-AE combinations that arise excessively often, i.e. with observed numbers greater than those expected [Suling and Pigeot, 2012].

Spontaneous Reporting databases

SR databases are large electronic databases comprising of systematically collected individual case safety reports for any suspected adverse drug reactions, which may come from a number of sources including medication manufacturers, consumers and healthcare providers. The WHO's Vigibase, for example, includes reports from centres in 60 countries, who are members of the WHO programme for international drug monitoring [Lindquist and Edwards, 2001]. Data-mining methods have been developed and applied in order to identify statistical associations in SR databases. Note that since the counts in such data come from spontaneously reported cases, exposures are based only on the frequency with which particular medications and AEs are reported together, rather than the true frequency at which they might occur in practice. Because SR databases only include records of individuals that were exposed to at least one medication and had at least one AE, there is also no healthy control or comparison group. Reporting rates and calculated relative reporting ratios are therefore relative only to that of other drugs and other AEs in the database, and cannot be generalised to the population of those who have not taken any medications and/or have not had any AEs that were suspected to be related to a medication. Reported associations from such analyses must therefore be regarded as hypotheses about possible relationships between the drugs and AEs, and not an approximation to the relative risk of the specific medication in relation to the CA in the general population. Of course (as with any other reported association from a statistical analysis), signals may also be caused by factors other than a causal relationship between the drug and AE in question (i.e. residual confounding). In order to confirm or refute a potential association and/or causality, any signals resulting from disproportionality analyses should be carefully followed up in further, more detailed investigations, epidemiological studies or randomized clinical trials [Gould A. Lawrence et al., 2015].

Multiple testing and frequentist methods of signal detection

The most widely used frequentist estimates of disproportionality are the Proportional Reporting Ratio (PRR) [Evans et al., 2001] and the Reporting Odds Ratio (ROR) [van Puijenbroek et al., 2002]. These measures are analogous to the relative risk and the odds ratio, but with the "exposed" individuals being those who have recorded exposure to the drug of interest and the "unexposed" those with no record of exposure to the particular drug of interest, but with exposure to at least one other drug in the database. That is, the relative frequency of spontaneous reports for a given drug and a specific AE (versus all other AEs) is divided by the corresponding quantity for all other drugs in the database or

study. This gives the PRR if frequencies are expressed as proportions and the ROR when they are expressed as odds. Other less commonly used frequentist approaches to signal detection include a cumulative sum method using cumulative numbers of drug-AE reports [Lao, 1997], a Poisson probability approach [Tubert et al., 1992], the use of propensity scores [Tatonetti et al., 2012] and large-scale logistic regression [Caster et al., 2010]. All these methods essentially produce a statistical measure of association for each drug-AE pair of interest, usually in the form of a score for which different thresholds can then be applied; any scores exceeding the chosen threshold then indicate an association between those particular drug and AE combinations. These estimators can be imprecise for rare drug-AE combinations (i.e. low cell counts), and the issue of multiplicity also arises due to the large numbers of drug-AE combinations of interest. This means that the overall type I error rate is likely to be inflated, which can lead to an unacceptably high number of false positive associations being identified as signals. Some frequentist methods have attempted to account for multiplicity using P-value adjustment; the double FDR procedure, for example, corrects P-values according to hierarchical groupings of codes [Mehrotra and Heyse, 2004], meaning that P-values are adjusted based on the number of events within a group of similar events, as opposed to simply adjusting for the number of events across all events. Another approach to reducing the number of false positive associations has been to increase the thresholds used to identify signals [Hauben and Reich, 2005, Slattery et al., 2013]. However, any choice of threshold is subjective, and represents a trade-off between missing potential signals (if the threshold is set too high) and creating too many false positives (if the threshold is too low) [Deshpande et al., 2010]. This trade-off between true detection rates and false positive rates is an important issue in surveillance programmes, because while it is essential for patient safety that true associations between drugs and adverse reactions are not overlooked, it is also important that resources are not wasted on signals that are likely to be false positives. Whilst adjustment for multiple testing is commonplace, minimising the false positive rate (type I error) for true null associations can therefore come at the cost of an increase in false negative rates (type II error). It has been advocated by some that adjustments for multiple comparisons should not be made at all [Rothman, 1990, Savitz and Olshan, 1995]. These authors highlighted some of the problematic implication of multiple testing adjustment, in particular that the interpretation of a test depends on whether or not other tests are conducted. Savitz and Olshan [1995], for example, suggested that not adjusting for multiple comparisons could avoid “unjustified dismissal of meaningful results or exaggerated confidence in weak results”.

Bayesian approaches to signal detection

Another approach to disproportionality analyses has been the widespread use of Bayesian shrinkage techniques. Shrinkage relates to the idea that an estimator may be improved if it is combined with other information, wherein the influence of all the estimates are considered simultaneously and hence "shrink" towards overall mean values. Estimates with large deviations from these overall means or those with greater uncertainty are penalised more strictly. Bayesian approaches to disproportionality analyses assume that reporting rates are similar (i.e. exchangeable) for all drug-AE combinations by assigning the combinations a common prior distribution. This can have a smoothing effect on the reporting rates estimated from the data. Such a prior may be estimated directly from the data at hand, estimated from previous data, or pre-specified by the researcher in a more subjective approach. If a minimally informative prior distribution is used then the smoothing effect is likely to be small, but estimates of disproportionality using Bayesian methods have generally been shown to shrink towards the null (in particular where there are low observed or expected counts) thus controlling for multiplicity by reducing the number of false positives (i.e. a conservative approach) [Roux et al., 2005]. However, this also means that positive associations may sometimes be less easy to discover and so the likelihood of false negative results can increase, once again highlighting the importance of the "trade-off" between detection and false positive rates. As in frequentist methods for signal detection, these Bayesian methods ultimately produce a statistical measure requiring some chosen threshold to identify drug-AE combinations as signals. The most commonly used Bayesian approaches to disproportionality analysis are the Gamma-Poisson Shrinker (GPS) [DuMouchel, 1999, DuMouchel and Pregibon, 2001] and the Bayesian confidence propagation neural network (BCPNN) [Bate et al., 1998]. These methods are used in practice to routinely detect signals for a number of large SR databases. Another Bayesian approach in this field has been the use of multi-level hierarchical Bayesian models, which have been applied to clinical trial data [Berry and Berry, 2004] and SR datasets [Crooks et al., 2012]. As well as incorporating Bayesian shrinkage, this use of hierarchical models is based on the idea that an estimation of the PRR (or ROR) of a drug-AE combination may be improved by combining this with information from other similar drugs (or AEs) according to specified groupings of these [Deshpande et al., 2010]. In this case shrinkage refers to the estimate for each member of a group being pulled in, (i.e. shrunk) towards the mean of all estimates in that group, as well as towards the null in general where there are small cell

counts and/or limited data. These three Bayesian approaches to disproportionality analysis are summarised briefly below.

The Gamma-Poisson Shrinker

The GPS is an empirical Bayesian data mining approach, where prior distributions used in the analysis are estimated from the data itself. The GPS produces empirical Bayesian geometric mean scores [DuMouchel, 1999, Fram et al., 2003], which are a measure of association similar to the PRR, such that a score greater than 1 implies an increase in the risk of a particular AE associated with a particular drug. The GPS model assumes that the observed count for any cell (i.e. any combination of a specific drug and specific outcome) follows a Poisson distribution. Since there are no denominator counts for the medication exposures in this type of data (with no unexposed cases), the expected counts are derived from the available data under the assumption that the exposures and the outcomes are independent. That is, the expected counts are calculated as the product of the marginal totals (the total number of records across the whole dataset for a particular exposure or for a particular outcome) divided by the total count of all observed records. The prior distribution used for the estimated disproportionality measure is a mixture of two gamma distributions, and each estimate is assumed to have a common prior distribution. To generate signals, cells are ranked using various approaches according to the posterior distribution of each PRR, for example the 5^{th} percentile is often used as a cut off [DuMouchel, 1999]. An example of the use of this method in practice is that of the US Food and Drug Administration, who use a GPS procedure for routine data mining system of their MedWatch database of voluntary reports of adverse drug events [Szarfman et al., 2002, Szarfman et al., 2004].

The Bayesian Confidence Propagation Neural Network

Bayesian neural network approaches to searching large numbers of drug-AE combinations are discussed in Bate et al. [1998]. The BCPNN aims to identify “unexpectedly” strong dependencies between drugs and AEs, as well as measuring how such dependencies change with the addition of new data. The measure of disproportionality used by the BCPNN is called the information component, which is a Bayesian implementation of the observed to expected ratio. A multinomial (rather than Poisson as in the GPS) model is used to produce shrinkage towards zero of the observed-to-expected number of AEs, with Bayesian prior parameters being fixed in advance (based on prior beliefs or existing knowledge/data) instead of estimated directly from the data. One example of the use of the BCPNN method

in practice is the World Health Organization programme for international drug monitoring, who have used this method since 1998 to identify drug safety signals in their international database of over two million case reports [Bate et al., 1998, Lindquist et al., 2000].

Bayesian Multi-level hierarchical models

A hierarchical structure based on AEs grouped within body systems has also been proposed in the form of a more flexible BHM, which was used to search for drug-AE signals in clinical trial data [Berry and Berry, 2004]. In the setting of a clinical trial, this model considers a three level hierarchy for the reporting of AEs and allows the generation of posterior distributions and risk differences. The comparison of the control and treatment groups is of less relevance to this thesis, but of note is the hierarchical structure that is implemented. This comprises three levels: the lowest level is the type of AE, the second level the body systems within which AEs could be grouped, and the highest level is then the collection of all body systems. In this framework of analysis, the priors for each drug-AE combination incorporate evidence from estimates for similar drugs and/or AEs in the same data. In addition, established expert epidemiological and medical knowledge can be incorporated. This model allows (rather than imposes) for the possibility that different AEs in the same body system might be related, and that rates of AEs are more likely to be similar within than across body systems. In these BHMs, the estimates for a particular drug-AE combination are adjusted towards the average of a defined group of similar drug-AE combinations, where members in the same group should be similar to each other in terms of relevant properties (e.g. drugs that act in a similar way or with a similar chemical makeup). The use of a mixture prior in the BHM [Berry and Berry, 2004] can allow for the possibility that many AEs could be completely unaffected by treatment by giving a point mass on the equality of the treatment and control rates. A mixture prior assigns some prior probability π to the expected proportion of null effects (i.e. where $PRR = 1$), and assigns a normal prior to the $(1 - \pi)\%$ thought to potentially have an effect (i.e. $PRR \neq 1$). A mixture prior for the $log(PRR)$ can then be expressed using a combination of two prior distributions for the proportion of expected null effects, where a stronger belief that there is no effect can be reflected by choosing a higher value of π. This model was further explored by Xia et al. [2011], who compared the use of normal and mixture priors for both binomial and Poisson BHMs and found that the use of a Poisson model gave statistical properties better suited for data with rare events. Crooks et al. [2012] further built on the model of Berry and Berry by considering confounding and interactions, application to SR data and specification of a more complex hierarchy. The model of Berry and Berry [2004] is

one-dimensional in that there is a single exposure variable (data come from the trial of a vaccine with treatment and a control groups) and so a hierarchical structure is only considered for the outcome variable (the AEs). An extension of this model that considered information sharing for groupings of medications and AEs simultaneously has been previously proposed [Brook, 2011]. Brook presented a theoretical formulation of this model by extending and combining the Gamma-Poisson Shrinker of DuMouchel [1999] and the hierarchical models of Berry and Berry [2004]. When this method was assessed using a sample of the WHO pharmacovigilance database, it was recommended that a two-dimensional model of information sharing could produce a more powerful BHM to detect true adverse drug reactions when compared to sharing information only in one dimension.

Comparisons of Frequentist and Bayesian approaches to signal detection

The different methods of disproportionality analysis described above have been evaluated and compared in a number of studies, with varying conclusions. One study, for example, compared six commonly used disproportionality measures (including those discussed above) when applied to the Netherlands Pharmacovigilance Foundation dataset, and found them to be largely comparable for combinations with at least four exposed cases [van Puijenbroek et al., 2002]. The authors recommended that a case-by-case approach be used when selecting a signal detection method according to the SR system or database in question. In another study, Xia et al. [2011] compared use of a BHM, a non-hierarchical Bayesian model, an unadjusted Fisher's exact test and two FDR procedures, and found that a BHM was helpful compared to other methods in reducing the number of false positives whilst improving the power to detect true signals. Candore et al. [2015] also evaluated five commonly used signal detection algorithms (PRR, ROR, GPS, BCPNN and a model based on Fisher's exact test) across three national or international SR databases as well as four safety databases from pharmaceutical companies, and found no method to be clearly superior but rather that the methods performed differently depending on which database was being considered. The authors therefore recommended that the absolute performance of a method should be assessed directly on the database of interest since its performance will be specific to that database [Candore et al., 2015]. In another study, Chen et al. [2015] compared eight methods for signal detection in terms of their detection and false positive rates, using data simulated to represent both SR databases and clinical trials. Here, BHMs were considered the most flexible approach, with consistently reasonable detection and false positive rates over a range of scenarios. On the other hand, the performance of BHMs was found to be unstable for small sample sizes (where the false positive rate increased)

and, compared to other methods, they were considered more complex and required longer computing times. Prieto-Merino et al. [2011] argued that the current standard methods of signal detection, e.g. those used in routine surveillance by the WHO and the FDA, do not exploit the full potential of Bayesian models. They recommend that the incorporation of medical knowledge and sensible hierarchies should be applied to order to share information across both medications and AEs in such databases.

4.2.3. Review of statistical methods for the detection of teratogenic medications in early pregnancy

SR databases such as the European Medicines Agency's EnduraVigilence [European Medicines Agency, 2016], the Uppsala Monitoring Centre's VigiBaseTM [Lindquist, 2008] and the US Foods and Drug Administration's AE reporting system [Sakaeda et al., 2013] include coding for CAs as a potential type of adverse drug reaction, and can therefore be used to try and identify teratogens. However, SR databases can be limited in their utility due to a number of factors, including duplicate case reports, reporting biases and, in particular, the potential for underreporting [Suling and Pigeot, 2012, Sharrar and Dieck, 2013]. Dissemination of a communication from the FDA regarding a drug's safety, for example, can affect the likelihood that an AE for a particular medication is reported [Ishiguro et al., 2014]. In general, if a medication has had attention in the media, individuals might be more likely to report on their use of it (i.e. recall bias). SR databases also tend to have limited coding in relation to CAs, for example they often do not distinguish between chromosomal and non-chromosomal CAs. Another limitation of SR databases in this respect is that they do not identify the timing of medication exposures, so it cannot be determined at which stage of development of the foetus a medication exposure occurred. Furthermore, SR databases do not include cases of termination of pregnancy for fetal anomaly. Population based CA registries, in contrast, use multiple sources to actively capture all cases of major CA (including terminations) in the population covered by each registry, thus minimising underreporting biases in terms of CA prevalence. Due to these limitations of SR databases other approaches to the detection of teratogens in CA data have been used. For example, patient registries have been used to compare the risk of major CAs following maternal use of specific medications. One such patient registry is EURAP, which collects data internationally on the use of antiepileptic medications during pregnancy [Tomson et al., 2011]. In these patient registries, women are registered before the outcome of their pregnancy is known, therefore minimising some of the biases encountered by SR databases. However, results from patient registries are limited to specific medications and

cannot therefore be used for signal detection across all types of medications. Instead, associations are typically evaluated in independent hypothesis-driven studies of specific types of medications or CAs. Carmichael et al. [2005], for example, performed a case-control study to assess whether maternal intake of progestin in early pregnancy was associated with an increased risk of hypospadias. Anderka et al. [2012] investigated the risk of selected CAs for medications used to treat nausea and vomiting during pregnancy. In another example, Zaqout et al. [2015] investigated the impact of the common first trimester dydrogesterone use on CHDs. Potentially informative relationships with other medications or other types of CA might be overlooked in these types of studies. Furthermore, studies of specific medications and CAs are not systematic or hypothesis generating and are therefore limited in their ability to identify new teratogens at the earliest possible stage.

Systematic signal detection for CAs in Europe: the EUROmediCAT network

The EUROmediCAT project was established in 2011 to build a European system for the evaluation of the safety of medication use during early pregnancy in relation to the risk of CAs [Morgan et al., 2011]. EUROmediCAT was built upon the existing network of EUROCAT, including only those registries with information on medication use during the first trimester of pregnancy [EUROCAT Central Registry, 2014b]. Only first trimester medication exposures are included in EUROmediCAT data because the critical period of development for most major CAs is in this period, during which time the organs of the foetus form [Czeizel, 2008].

Current EUROmediCAT signal detection methodology

A hypothesis-generating signal detection method has recently been developed, which uses the EUROmediCAT database to routinely identify potential teratogenic medications taken during the first trimester of pregnancy [Luteijn et al., 2016]. This method uses a one-sided Fisher's exact test to compare the odds of exposure to a specific CA and medication to the odds of exposure to the same medication in the remainder of the dataset, i.e. all other medication-exposed CAs. Each medication and each CA is examined separately, using ATC codes for medication exposures (see Table 1.1) and EUROCAT subgroup codes for CAs (see chapter 3). Only medications coded to ATC level four (ATC4) or five (ATC5) are considered, giving five or seven digit codes, respectively, containing information regarding the chemical subgroup and the chemical substance of a medication. Exposures with information at only ATC3 or below (i.e. four or less digits per code) are excluded, as they do not provide precise

enough detail for the purposes of signal detection analyses. A separate analysis is performed for each medication-CA combination. An FDR procedure is then applied (see next section for further details) in order to adjust for multiple testing when determining the statistical significance of each test. A separate analysis is performed for ATC4 and ATC5 codes, with duplicate statistically significant associations being excluded from the set of potential signals. Duplicate associations here refers to those involving an ATC4 code where a more detailed ATC5 code is associated with the same CA (e.g. if N03AG and N03AG01 were both signals with the same CA), or those involving aggregate CA codes where a more specific code was associated with the same medication (e.g. if NTDs and spina bifida were both associated with the same ATC code). Statistically significant associations showing a protective association are not further investigated, since these are thought to be likely due to chance or due to bias arising from the study design [Luteijn et al., 2016]. A protective association is only in comparison to other CAs and medications in the database as there are no "healthy" controls, and does not imply that any particular medication is associated with a lower overall risk of a particular CA. This signal detection methodology was applied to EUROmediCAT data for the years 1995–2011, and it picked up some (but not all) known teratogens, as well as identifying new potential associations. The use of different FDR cut-offs were assessed and an FDR of 50% was found to provide a reasonable balance between detection rate and minimising the workload created in terms of having to follow up potential signals; this means that up to 50% of medication-CA combinations found to be potential signals are expected to be false positive associations. A total of 39 combinations were considered signals after the FDR procedure, 28 of which were for antiepileptic, antidiabetic, antiasthmatic medications or selective serotonin reuptake inhibitors; these medication groups are already examined separately as part of other EUROmediCAT projects [de Jong-van den Berg et al., 2011] and were therefore not considered further in the signal detection process. The remaining 11 signals were discussed in a separate paper that examined the potential new associations in detail [Given et al., 2016]. In this paper an additional 16 signals were also considered for further examination, based on a previous signal detection analysis of the same EUROmediCAT dataset that had been reported prior to analytical refinements such as the combination of duplicate ATC codes and the specific (single FDR-adjusted) cut-off P-values for associations in ATC4 and ATC5 coding [EUROmediCAT, 2015]. Exposed cases for medication-CA associations identified using the signal detection method were then validated with local registries to confirm diagnoses and the timing and type of medication exposures. Odds ratios based on the validated data were

then adjusted for confounding by registry, and medication-CA associations persisting after this process were considered to be validated statistical signals. After data validation, Given et al. [2016] found that there remained evidence for a signal in 13 of the initial 27 associations considered, for which a literature review was then performed to assess existing evidence of human teratogenicity. Prior evidence was found to support 6 of the 13 signals, with the other signals requiring further confirmation in an independent dataset.

Use of a false discovery rate procedure to adjust for multiple tests

When determining the significance of each test separately, the conventional cut-off level of 5% for statistical significance means that 5% of all medication-CA combinations with a statistically significant result will be labelled as signals due to chance alone. For example, if 1000 tests are performed then 50 of these are expected to be statistically significant, but not true associations. It can then be difficult to determine a true association (if indeed there are any) amongst all of the positive results, due to the proportion that that are not true associations; if there was only one true association in this example it would need to be picked out from the 51 statistically significant results. One approach to this problem is to control the proportion of false positive results among the set of all positive results; this is the FDR. In other words, the FDR is the proportion of incorrect rejections amongst all rejections of the null hypothesis. FDR control can be achieved using a multiple testing method such as the Simes procedure [Benjamini and Hochberg, 1995]. In the above example the FDR would be 50/51, implying that around 98% of identified signals might be false positive associations.

Limitations of current EUROmediCAT methodology and rationale for methodology applied in this thesis

A main strength of the EUROmediCAT signal detection method is that it is systematic and ongoing, in that it aims to be repeated when new data becomes available. The issue of multiple testing is also addressed through use of an FDR adjustment. However, this adjustment for multiple testing is done across the whole database and, as discussed in Chapter 1, potential relationships between medications or CAs are not considered. Medications in the same ATC classes are often known to work in similar ways, and this information may be useful for signal detection methodology. Similarly, certain CAs are thought to be more sensitive (compared to other types of CAs) to medications in general. Methods used in pharmacovigilance for SR and clinical trial data are able to specify that medications in the same class (e.g. within a particular chemical or therapeutic subgroup) are expected to have similar teratogenic properties; instead of one association at a time,

associations for groups of related medications can be considered simultaneously. These methods have not yet been explored for use in routine single detection for population-based CA data. An objective of this thesis was therefore to refine the signal detection component of EUROmediCAT surveillance methodology. To achieve this, two different approaches were considered

1. using post-analysis FDR adjustments that take groups of medications and/or CAs into account (Chapter 5)
2. using BHMs to directly model potential group effects for groups of similar medications and/or CAs (Chapter 6)

The statistical models used in the following two chapters incorporate aspects of the current EUROmediCAT methodology combined with signal detection methodology used in more general pharmacovigilance settings (i.e. SR databases or clinical trials data), such that models applied in this thesis are novel approaches to the routine signal detection analyses of population-based CA data.

Calculating measures of disproportionality congenital anomaly data: the "exposed malformed" design

The "exposed malformed" design refers to the fact that all individuals in the dataset are CA cases that have been exposed to at least one medication; as such, the "controls" in each comparison are malformed foetuses with at least one major CA and who were exposed to at least one medication *other than* the specific CA and medication under consideration. Table 4.1 displays an example 2x2 table for the test of association between a specific medication and a specific CA, for a measure of the risk associated with medication i for CA j, compared to all other CAs and medications in the data.

Table 4.1. The "exposed malformed" design in analysis of the relationship between a medication i and a congenital anomaly (CA) j.

	Cases: Foetuses with CA j	**Malformed controls:** Foetuses with CAs other than j	Total
Exposed to medication i	c_{ij}	$c_{ij'}$	$c_{i.}$
Unexposed to i, but exposed to at least one other medication in the data	$c_{i'j}$	$c_{i'j'}$	$c_{i'.}$
Total	$c_{.j}$	$c_{.j'}$	$N\ (= c_{..})$

The most commonly used measures of disproportionality are the ROR and the PRR. These can be explained in the context of EUROmediCAT data using notation from Table 4.1, which presents the observed count c_{ij} for the number of exposures to drug i for foetuses with CA j. The ROR and PRR are defined as follows

$$ROR_{ij} = \frac{c_{ij}/c_{ij'}}{c_{i'j}/\ c_{i'j'}}$$

$$PRR_{ij} = \frac{c_{ij}\ /\ c_{i.}}{c_{i'j}\ /\ \ c_{i'.}}$$

A PRR or ROR of 1 indicates that there is no suspected association between the medication and CA of interest.

Use of the proportional reporting ratio

The PRR was the measure used for signal detection analyses in this thesis, as this is naturally used by models for count data. RORs and PRRs have been shown to be similarly effective in practice as measures of disproportionality [van Puijenbroek et al., 2002, Waller et al., 2004]. The link between these two measures is similar to the approximation of the rate ratio through the use of an odds ratio in a case-control study, and the use of the PRR rather than the ROR should not have a material effect on our conclusions. Indeed, the two measures are known to give very similar results when the count c_{ij} is a low proportion of the total exposures to medication i $(c_{i.})$, and the count $c_{i'j}$ is a low proportion of the total exposures to all other medications in the database $(c_{i'})$. This may not be the case for all types of CA; the CHDs, for example, are the most common type of CA and these affect up to 35% of cases in the EUROmediCAT data [Luteijn et al., 2016]. The counts c_{ij} and $c_{i'j}$ are therefore likely to represent a relatively high proportion of the total exposures in the dataset, especially for some of the more common CHD subgroups.

Protective associations in signal detection analyses

Statistically significant associations resulting from models that showed a protective association (i.e. a PRR < 1) were not flagged as signals, as is the case in the existing EUROmediCAT methodology. The aim of signal detection is to identify potentially harmful medications, and this is why a one-sided significance test is used in EUROmediCAT analyses. This does not mean, however, that no protective associations are expected; some are likely to occur e.g. due to chance or bias arising from the study design. The number of combinations showing a protective association were therefore monitored throughout all models applied, with potential causes for these associations being considered. Note that a

protective associations here does not imply that a particular medication is associated with a lower overall risk of a particular CA, since there are no healthy controls and hence the comparison is only to other CAs and medications in the database.

4.3. Validation and comparison of signal detection methods for CA data

A way of judging how good these different methods are was required in order to be able to decide which method is most appropriate for use in the analysis of medication safety during pregnancy. For example, it would be useful to be able to quantify how many medications that are known to be harmful are being picked up by each method. Aside from a few well-known exceptions, however, there is in fact very little available or substantial evidence regarding which medications are likely to be teratogenic. This was highlighted by a recent comprehensive review, which confirmed a lack of existing knowledge on the teratogenic effect of medicines used during pregnancy [van Gelder et al., 2014]. This review emphasised that prescription rates are not associated with current knowledge on teratogenicity of many medications; those known to be most commonly used during pregnancy were not generally the medications for which teratogenic risks have been well studied or evaluated. Nevertheless, associations reported from case-control studies in this review were used to evaluate the signals obtained by the recently developed EUROmediCAT signal detection system, resulting in only a small set of eight associations available for validation [Luteijn et al., 2016]. In order to consider the relative value of the methods quantitatively, however, a more comprehensive set of known medication-CA associations would ideally be identified. The proportion of associations detected by each method could then be calculated and directly compared.

4.3.1. Risk classification systems for the prescription of medications during pregnancy

The Australian classification system for prescribing medicines in pregnancy [Australian Government Department of Health, 2016; https://www.tga.gov.au/prescribing-medicines-pregnancy-database] was identified as a potential source of information for use as a comprehensive means of method validation, where the number of “high risk” medications (as independently identified using this database) could be used to quantitatively judge the signal detection methods for CA data. This system was established in 1963 by the Australian Medication Evaluation Committee, to advise on the safety of new medications being introduced into Australia and to monitor and evaluate potential AEs of medications already in use. This provides a database for use when prescribing medicines in pregnancy, which is

developed and maintained by medical and scientific experts according to available evidence of recorded risks associated with taking particular medicines while pregnant. All medications in this database are divided into five main lettered categories, which are summarised in Table 4.2. Category A medications are considered to be safe for use during pregnancy. Medications in categories B1, B2 and B3 have not shown any evidence of harmful effects or increased frequency of malformations for human foetuses. Those in category C may carry harmful effects to human foetuses, but without any evidence of causing malformations. Finally, medications in categories D and X are considered to carry moderate to high risk as they are believed to increase the frequency of human fetal malformations and can lead to permanent and irreversible damage.

Table 4.2. Definition of categories in the Australian system for prescribing medicines in pregnancy.

Category	Summary of medications in category	Risk [a]
A	• Taken by large number of pregnant women /women of childbearing age • No increase in frequency of malformations observed	Low
B1-B3	• Taken by limited number of pregnant women /women of childbearing age • No increase in frequency of malformations observed B1: no evidence of increased fetal damage from animal studies B2: no evidence of increased fetal damage from animal studies, available studies inadequate or lacking B3: evidence of increased fetal damage in animal studies, significance for humans uncertain	Low
C	• Caused /suspected of causing harmful effects on human foetus, without causing malformations • Effects may be reversible	Low
D	• Caused/ suspected /expected to cause increased incidence of human fetal malformations, or irreversible damage	High
X	• Medications with a high risk of causing permanent damage to the foetus • Should not be used in pregnancy or if there is a possibility of pregnancy	High

[a] Risk associated with causing malformation to the foetus, categorisation as defined for use in this thesis

A teratogenic signal detection method should be expected to pick up more medications in the "high risk" categories D and X than in the "low risk" categories A-C. As the proportion of

medications in categories D and X that are identified as signals increases this should indicate improvements in the sensitivity of a signal detection method.

The Australian categorisation system gives an online table of all medications in its database with five fields of information for each medication listed. These fields are medication name (given by the active ingredient or the generic medication name rather than any trade or commercial names), specified risk category and three levels of classification according to the pharmacological group or action of the medication. Further additional information can be obtained from the database for certain medications, including a full description of the given risk category and any safety statements relating to the particular medication, where applicable. An example of the information given for one such medication is displayed in Table 4.3.

Table 4.3. Example of information given for a specific medication in the Australian prescribing medicines in pregnancy database (taken from https://www.tga.gov.au/prescribing-medicines-pregnancy-database).

Name:	Aliskiren
Category:	D
Category Description:	Drugs which have caused, are suspected to have caused or may be expected to cause, an increased incidence of human fetal malformations or irreversible damage. These drugs may also have adverse pharmacological effects. Accompanying texts should be consulted for further details.
Classification 1:	Cardiovascular System
Classification 2:	Antihypertensive
Classification 3:	Angiotensin II receptor antagonists and renin inhibitors
Additional Information:	When used in pregnancy during the second and third trimesters, drugs that act directly on the renin-angiotensin system can cause injury and even death in the developing foetus. Although no adverse fetal effects have been linked to first trimester drug use of ARAs, the number of exposures reported is too small to determine conclusively that ARAs are safe in the first trimester. Pregnant women who are taking ARAs should be changed as quickly as possible to other antihypertensive medication to maintain normal blood pressure. It is generally advisable not to use ARAs for the management of hypertension in women who are likely to become pregnant.

It can be seen in this example that the additional information (for which many substances do not have anywhere near this amount of detail, if any at all) contains little information specific to particular CAs. For this particular medication, the risks to the foetus at different trimesters of pregnancy is unclear and although it is classified as a “high risk” medication,

there is inconclusive evidence about the risks in early pregnancy, so may not in reality be a high risk medication for first trimester exposure data.

Issues with the use of risk categorisation systems to address the safety of medications during pregnancy

Similar categorisation systems have been in place in Sweden since 1978 and the US since 1979, including pregnancy labelling regulations and the introduction of similar letter risk categories. However, the distribution of drugs into the various categories is sometimes known to vary between the Swedish, US and Australian systems [Sannerstedt et al., 1996]. There are also further issues with these types of categorisation systems. There has not been sufficient research into the effects of many medications, for example, to be able to adequately assess their risks for human foetuses, and consequently medications within the same risk category do not necessarily have similar risks. This issue is particularly relevant to the US categorisation system, where the majority of medications have generally been allocated to the "risk cannot be ruled out" category due to the stringent quality of data required to classify a medication as being "high risk" [Sannerstedt et al., 1996]. This issue contributed to the decision to introduced new labelling rules to replace the food and drug administration's letter risk categorization system in the US in June 2015 [US Food and Drug Administration, 2014]. The updated system requires labelling for new medications to include three detailed subsections describing the risks of that medication during pregnancy, not only in the context of fetal risk but also other factors including maternal disease severity, co-existing conditions and potential alternative therapies [Mosley et al., 2015]. New medications in the US are therefore no longer assigned to one of the letter risk categories. These new rules aimed to provide more detailed information for each individual medication and to better guide clinical decisions for pregnant and breastfeeding women. Other concerns about this type of classification for medications include a lack of differentiation between uses of the same medications for different conditions, and the fact that the stage of pregnancy at which the medications might be taken is not considered [Ramoz and Patel-Shori, 2014]. In particular for analyses of CA data, it is important to note that there is no information as to which specific CAs are affected by any particular medication (e.g. as can be seen in Table 4.3). As such, whilst classifications are made separately for medications they are not specific to different types of CA. If a medication is assigned to the "high risk" category this is the case for all combinations of that medication with a CA, despite the fact that the medication may only be teratogenic for certain types of CA in reality. The antiepileptic medication valproic acid, for example, is a well-known

teratogen that has been associated with an increased risk of certain CAs, including spina bifida, atrial septal defect, cleft palate, hypospadias, polydactyly and craniosynostosis [Jentink et al., 2010]. There is no evidence, however, that valproic acid is also a "high risk" medication for all other CAs, so we may not expect it to have any effect on the relative PRR of all other types of CA. The set of all medication-CA combinations that are identified as being "high risk" in EUROmediCAT data is therefore likely to be overestimated in this thesis. A further complication specific to these data is that medications in category X are likely to be underrepresented in the EUROmediCAT data since pregnant women are advised against these known teratogens. In fact, as soon as a medication becomes known as a teratogen, there can be a rapid switch in prescription practises worldwide. Some exposures to such medications would still be expected in CA databases e.g. due to medication use during early stages of unplanned pregnancies, or for women who might be taking a medication for a chronic or life threatening condition. However, there is likely to be insufficient numbers to provide the power to detect an association in such cases, even when the medication is known to be truly teratogenic.

Lack of a gold standard reference set of known teratogens

Due to the above issues, none of these classification systems can be considered as a gold standard against which the results from the current signal detection methods may be compared and evaluated. However, what was required for this thesis was a way of directly comparing the different possible methods of signal detection that may be used in practice with CA data. Whilst the categorisation system cannot provide an absolute measure of how good any chosen model is, each model will have the same lack of data and power for known teratogens, so use of measures based on the Australian categorisation system were considered a potentially useful way of directly comparing the relative strengths of the models applied to EUROmediCAT data in the following chapters.

Mapping the Australian risk categories to EUROmediCAT data

In order to compare different methods using the proportion of signals detected in "low" and "high" risk categories (see Table 4.2), the online database of the Australia categorisation system was downloaded in table format [Australian Government Department of Health, 2016]. Medication names given in this database were matched to the EUROmediCAT data using substance names for ATC coded medications. Where mismatches occurred, edits to the medication names in each database were made

manually. Where available, the given risk category according to the Australian classification system was then included for each specific ATC code in the EUROmediCAT data.

4.3.2. Defining measures for the comparison of signal detection methods

In signal detection analyses, the detection rate is the proportion of all the true positive associations that are identified as signals. Of course, it is not possible to get a realistic estimate of this detection rate for CA data since there is no comprehensive or reliable reference list of all true associations. For this thesis, therefore, the proportion of all the "high risk" medications that were identified as signals was calculated using the risk categories of the Australian classification system. This was called the identification rate, and was defined as

$$\textbf{Identification rate} = \frac{\text{Number of "high risk" medications identified as signals}}{\text{Total number of "high risk" medications in the data}}$$

The proportion of medication signals that were "high risk" out of all the identified medication signals (i.e. including those with "low" or unidentified risk) was also be used to compare the different methodologies considered. This can be thought of as an estimate similar to the positive predictive value, which is the number of true positive associations out of the total number of associations identified as signals. For this thesis, the "high risk" proportion was defined as

$$\text{"}\textbf{High risk}\text{" }\textbf{proportion} = \frac{\text{Number of medication signals in "high risk" category}}{\text{Total number of medications identified as signals}}$$

Another important consideration in signal detection analyses is to balance the potential workload with the false discovery and detection rates. Every association flagged as a potential signal needs to be followed up separately in further studies, which requires time and effort for each individual signal being assessed. It is important, therefore, that whilst achieving the highest possible detection rate, a realistic restriction is placed on the total number of signals identified that can be followed up in practice. This must also be balanced with the number of signals that are likely to be true associations such that time and resources are not wasted in following up large numbers of false positive associations. Alongside the identification rate and "high risk" proportion, methods were therefore also compared in terms of how many signals they identified, in order to quantify the workload created by each method. This is referred to as the "effective workload" and was defined as follows

$$\textbf{Effective workload} = \text{Total number of medication signals}$$

Note that rather than the total number of medication-CA combinations identified as signals, the number of medications within this set of signals is considered the resulting workload. This reflects the fact that the actual medications are considered the risk factors, whether they are associated with one or more CAs. In terms of the required follow up of potential signals, each additional medication flagged up as a signal presents a more significant increase in workload than any number of additional signals of a medication that has already been flagged (with a different CA).

4.4. EUROmediCAT data

4.4.1. Data sources and included congenital anomalies and medications

Data for these analyses were from 13 EUROmediCAT registries in 11 countries that agreed to participate in this study, covering a total population of around 7 million births in the period 1995-2011. Note that this did not include two of the registries (Mainz; Cork and Kerry) that contributed data to the previous EUROmediCAT signal detection analyses, and for two registries (Zagreb; Poland Weilkopolska) an additional year of data from 2011 was available that had not been included previously. As the exact same dataset could not be obtained for this project, results using the current EUROmediCAT methodology would not be identical to those published by Luteijn et al. [2016]. However, since the majority of the data used was the same, any signals detected or overall conclusions drawn when using the same single FDR methodology were considered comparable.

Congenital anomalies monitored for the purposes of signal detection

EUROmediCAT data were coded in the same way as the data on CAs in EUROCAT (see Chapter 3), here consisting of cases of non-genetic CAs. Cases with any of the following CAs were excluded, since they are inherited and therefore ca not potentially be caused by teratogenic medications: chromosomal anomalies, skeletal dysplasia, genetic syndromes and microdeletions. Foetuses with isolated congenital dislocation of the hip as their only major CA were also excluded, since the aetiology of this CA is known to be mechanical (i.e. caused by physical pressures from outside the uterine environment). The cases in the analysis were foetuses with that specific CA for 55 pre-defined EUROCAT subgroups [EUROCAT Central Registry, 2013]. The highest level of EUROCAT coding gives the major organ subgroup, within which there are further classes; these higher level groups are also referred to as aggregate subgroups. The aggregate subgroups include all cases listed in the subgroups at lower levels as well as cases that do not have this more detailed information. For example, if a foetus is recorded as having a CHD, but the particular CHD subgroup is not

recorded, this case will still be counted in the aggregate CHD subgroup. In the previous EUROmediCAT signal detection analyses, the only aggregate subgroups that were monitored were NTDs, CHDs and severe CHD. In EUROmediCAT data, the vast majority of cases in aggregate CA groups are also attributed a more specific subgroup code, so including these groups when considering CAs simultaneously means that a large number of cases would be counted twice in the analysis. In the analyses of Luteijn et al. [2016], this issue was dealt with at a later stage by removing duplicate statistically significant associations involving aggregate CA codes where a more specific code was associated with the same CA. For this thesis, however, no aggregate subgroups were monitored because this would have led to a large overlap of information when considering methods that grouped medication--CA combinations. For the aggregate group of NTDs, all the cases in this dataset also had a more specific code giving the type of NTD. For those with a recorded CHD, however, 542 had no further information regarding the specific type of defect; these cases were therefore combined into a separate subgroup ("Unspecified CHD") for analysis purposes. In addition, two subgroups that were monitored in the recent EUROmediCAT signal detection analysis (neural crest and complete absence of a limb) were not included here because a more recent version of EUROCAT coding was used for the extraction of data for this thesis. The newer version of coding also included five further CAs that were not part of the previous signal detection analyses and hence (for comparison purposes) these were not monitored for this thesis. The following EUROCAT subgroups were included in the data as malformed controls, but are not monitored for signal detection: cystic adenomatous malformation of lung, hypoplastic left heart, indeterminate sex, congenital skin disorders, teratogenic syndromes with malformations, fetal alcohol syndrome, valproate syndrome, and maternal infections resulting in malformations.

Medication exposures in EUROmediCAT data

Maternal medication exposure data in EUROmediCAT are typically obtained from prospectively recorded maternity records [Boyd et al., 2011, Bakker and Jonge, 2014]. Information on medication exposures are collected by registries, with some registries also having additional data sources such as general practitioner records, maternity passports, maternal interview before or after birth and medical records of the infant [Bakker and Jonge, 2014]. Inclusion in EUROmediCAT database requires exposures to have occurred in the first trimester, which is defined as the time from the first day of a woman's last menstrual period up to her twelfth week of gestation. The two Polish registries, however, were known to include exposures outside of this time and so data cleaning by time of

exposures was performed to ensure that only exposures that were known to have occurred during the first trimester were included. EUROmediCAT codes all medications using the ATC system (see Table 1.1), with an unlimited possible number of first trimester exposures being coded, each including up to seven digit codes and free text information. Malformed foetuses that were exposed to at least one recorded medication in the first trimester of pregnancy were included as cases in the data for analyses. Foetuses that were exposed only to vitamins, minerals and/or folic acid were excluded. Cases exposed exclusively to medication codes with less than 5 digits (i.e. not coded up to at least ATC4 level), topical medications (ATC codes S01-S03, D01A, D02-D04, D05A, D06-D09, D10A, D11AA, D11AC, D11AE, D11AF, D11AH01-D11AH03, M02 and all D11AX codes except for oral preparations) and those taken in the 2nd/3rd trimester or with unknown timing were also excluded. ATC codes subject to alterations over time were obtained from the WHO website [WHO Collaborating Centre for Drug Statistics Methodology, 2015], and older codes were updated where available. ATC5 codes were analysed; where there was only information available to ATC4 level, this was included in the analysis as a separate code. As the majority of codes in the dataset included information up to ATC5 level, this allowed these few ATC4 codes to be incorporated in the groupings. This meant as much information was included as possible whilst avoiding duplication of analyses and results that occur when ATC4 and ATC5 codes are analysed separately.

4.4.2. EUROmediCAT data description

A summary of the EUROmediCAT dataset as extracted for this thesis is displayed in Table 4.4. A total of 31,197 foetuses with at least one first trimester medication exposure (excluding exposures to only folic acid, minerals and/or vitamins) and a major CA (excluding genetic conditions) born from 1995 to 2011 were extracted from the EUROmediCAT central database for 13 registries. Of these, 905 foetuses with isolated congenital dislocation of the hip, 1,219 with no ATC4 or ATC5 level medication exposures recorded and 452 with only topical medication exposures were excluded, leaving 28,621 foetuses with valid medication exposures. Foetuses with exposures outside the first trimester of pregnancy (n=1,490) or with unknown timing (n=12,073) were further excluded, for a remaining 15,058 foetuses available for analysis. Poland and Wielkopolska registries had the largest data loss due to unknown exposure timings, with over 82% of foetuses being excluded where it was not possible to verify whether the medications reported had in fact been taken in the first trimester of pregnancy. The next highest proportion of data loss due to unknown exposure

timings was for the Northern Netherlands registry, where notes in records about timings led to the exclusion of 25% of records. However, the distribution of types of CA were similar for those pregnancies included and excluded due to unknown timing, suggesting that the cases remaining in the dataset for these registries should not be prone to selection biases in this respect. All other registries had less than 5% data loss due to cleaning by timing of medication exposures. On average, there were 1.55 recorded ATC-coded non-topical medications per pregnancy, ranging from one (in 65% of cases) up to 16 (in one case) recorded medication exposures per pregnancy. The total number of medication exposures for the 15,058 foetuses was 23,410. Of these, 22,624 were exposures to medications that appeared at least 3 times in the dataset. Around 4% of exposures were coded using only a five digit ATC4 code (n=1,037) rather than a full seven digit ATC5 code. Overall, there were 893 unique ATC5 medication codes in the data, and a further 123 codes with information only up to ATC4. After the exclusion of medication codes with less than 3 exposures overall, 523 ATC medications remained for analyses, of which 39 (7.5%) were coded only to ATC4 level. With 55 CAs, this gave a total of 28,765 potential medication-CA combinations available for analysis.

Table 4.4. Description of data from 13 EUROmediCAT registries for the analysis of safety of medication use during first trimester of pregnancy.

EUROCAT Registry	Birth years included	Foetuses with CAs and at least one valid exposure	Foetuses with CAs following data cleaning by timing of exposure [a]	Data loss by data cleaning (%)	Total eligible ATC coded exposures	Total ATC codes excluding those with <3 exposures	Average ATC coded medication exposures per pregnancy
Denmark, Odense	1995-2011	240	240	0	367	346	1.53
France, Paris	2001-2011	658	658	0	970	897	1.47
Italy, Tuscany	1995-2011	1,083	1,033	4	1,417	1,352	1.37
Netherlands, North Netherlands	1995-2011	2,451	1,848	25	3,133	2,933	1.70
Italy, Emilia Romagna [b, c]	1995-2011	2,350	2,349	0	3,860	3,736	1.64
Switzerland, Vaud	1997-2011	309	297	1	458	433	1.54
Croatia, Zagreb	1995-2011	198	190	2	243	218	1.28
Malta	1996-2011	306	305	0	461	453	1.51
Belgium, Antwerp	1997-2011	349	347	1	508	478	1.46
UK, Wales	1998-2011	2,057	2,057	0	3,030	2,924	1.47
Norway	2005-2010	3,051	3,051	0	5,535	5,481	1.81
Poland, Wielkopolska	1999-2011	3,180	469	85	640	632	1.36
Poland (excluding Wielkopolska)	1999-2010	12,389	2,214	82	2,788	2,741	1.26
Total	1995-2011	28,621	15,058	47	23,410	22,624	1.55

[a] After exclusion of CA registrations with only medication exposures of unknown timing

[b] During the period 1995 to 2004 Emilia Romagna database had space for only 5 medications to be recorded

[c] Terminations of pregnancy for fetal anomaly were excluded from the Emilia Romagna registry as information on medications is only available for live and still births

Table 4.5 displays the number of foetuses and the proportion (of 15,058 malformed foetuses) affected by each type of CA. Over a third of malformed foetuses exposed to at least one first trimester medication were born with a CHD, the most common type of which was ventricular septal defect (17% of all exposed cases).

Table 4.5. Number of cases with a congenital anomaly (n=55) analysed for signal detection in 15,058 malformed foetuses.

Type of CA	Congenital Anomaly [a]	Index [b]	Malformed foetuses	
			N	%
Nervous System	**Neural tube defects**	-	**562**	**3.73**
	Anencephalus	1	162	1.08
	Encephalocele	2	88	0.58
	Spina Bifida	3	312	2.07
	Arhinencephaly/holoprosencephaly	4	44	0.29
	Hydrocephaly	5	308	2.05
	Microcephaly	6	121	0.80
Eye	Anophthalmos/microphthalmos	7	79	0.52
	Congenital cataract	8	87	0.58
	Congenital glaucoma	9	28	0.19
Ear, face & neck	Anotia	10	23	0.15
Heart	**Congenital heart defects (CHDs)**	-	**5,250**	**34.86**
	Severe CHDs	-	**1,300**	**8.63**
	Aortic valve atresia/stenosis	11	122	0.81
	Atrioventricular septal defect	12	144	0.96
	Coarctation of aorta	13	226	1.50
	Common arterial truncus	14	38	0.25
	Ebstein's anomaly	15	32	0.21
	Hypoplastic right heart	16	22	0.15
	Pulmonary valve atresia	17	73	0.48
	Single ventricle	18	56	0.37
	Tetralogy of Fallot	19	207	1.37
	Total anomalous pulmonary venous return	20	31	0.21
	Transposition of great vessels	21	234	1.55
	Tricuspid atresia and stenosis	22	56	0.37
	Atrial septal defect	23	1,342	8.91
	Pulmonary valve stenosis	24	308	2.05
	Ventricular septal defect	25	2596	17.24
	Patent ductus arteriosus (only CHD in term infants)	26	264	1.75
	Unspecified CHD	27	542	3.60

[a] *aggregate subgroup codes are not included in signal detection analyses for this thesis (in bold)*

[b] *This index identifies the congenital anomalies in Figure 4.1 and Figure 6.5–Figure 6.9*

The next most commonly occurring CAs were atrial septal defects and hypospadias, with over 8% of malformed foetuses affected with each of these. The aggregate subgroups of NTDs, CHD and severe CHDs are not included in signal detection analyses as separate subgroups due to their overlap with the more specific subgroups.

Table 4.5 (continued). Number of cases with a congenital anomaly (n=55) analysed for the purpose of signal detection in 15,058 malformed foetuses.

Type of CA	Congenital Anomaly [a]	Index [b]	Malformed foetuses	
			N	%
Respiratory	Choanal atresia	28	38	0.25
Oro- facial clefts	Cleft lip ± palate	29	713	4.73
	Cleft palate	30	503	3.34
Digestive system	Ano-rectal atresia and stenosis	31	222	1.47
	Annular pancreas	32	16	0.11
	Atresia of the bile ducts	33	17	0.11
	Atresia/stenosis, other parts of small intestine	34	61	0.41
	Diaphragmatic hernia	35	188	1.25
	Duodenal atresia or stenosis	36	58	0.39
	Hirschprung's disease	37	61	0.41
	Oesophageal atresia	38	182	1.21
Abdominal wall	Gastroschisis	39	167	1.11
	Omphalocele	40	139	0.92
Urinary	Bilateral renal agenesis	41	71	0.47
	Bladder exstrophy and/or epispadia	42	44	0.29
	Congenital hydronephrosis	43	788	5.23
	Multicystic Renal dysplasia	44	209	1.39
	Posterior urethral valve and/or prune belly	45	73	0.48
Genital	Hypospadias	46	1,291	8.57 [c]
Limb	Club foot	48	848	5.63
	Limb reduction defects	47	443	2.94
	Polydactyly	49	613	4.07
	Syndactyly	50	386	2.56
Other anomalies /syndromes	Congenital construction bands	52	33	0.22
	Conjoined twins	53	4	0.03
	Craniosynostosis	51	132	0.88
	Laterality Defects [d]	54	82	0.54
	Situs inversus	55	53	0.35

[a] *aggregate subgroup codes are not included in signal detection analyses for this thesis (in bold)*

[b] *This index identifies the congenital anomalies in Figure 4.1 and Figure 6.5–Figure 6.9*

[c] *Hypospadias are a birth defect of the urethra in males only, therefore affecting 14.75% of male foetuses*

[d] *The laterality defects subgroup comprises CAs where an organ has formed on the wrong side of the body, including: atrial isomerism, dextrocardia, situs inversus, broncho-pulmonary isomerism, asplenia and polysplenia.*

Table 4.6 shows the number of EUROCAT coded CAs per malformed foetus in 15,058 women. The average number of CAs per malformed foetus was 1.3, and 79% of foetuses were affected with only one recorded major CA. When considering the 55 CAs monitored for the purposes of signal detection in this thesis, 85% of foetuses were affected with only one CA and there was an average of 1.2 CAs recorded per pregnancy. Fifteen percent (n=1,887) of malformed foetuses in the dataset had more than one of the 55 monitored CAs recorded, with a maximum of seven CAs recorded in one particular foetus.

Table 4.6. Number of congenital anomalies (CAs) per malformed foetus for 15,058 pregnancies in signal detection dataset.

Number of CAs per malformed foetus	**Number of pregnancies: counts for all CAs in data [a]**	**Number of pregnancies: counts for 55 CAs included in signal detection analyses**
1	11,857	10,524
2	2,292	1,431
3	571	339
4	210	81
5	82	29
6	33	6
7	11	1
8	2	0
Total	15,058	12,411 [b]
Average CAs per malformed foetus	1.3	1.2

[a] The counts are for all CAs excluding congenital dislocation of the hip and genetic conditions
[b] 2,647 individuals only had CAs that are not monitored for signal detection purposes. These are included in the data as malformed controls.

Table 4.7 shows the number of first trimester medication exposures for 15,058 malformed foetuses according to each anatomical main group (ATC1), with counts for the most common ATC2 groups. The most common type of medications used in the first trimester were genitourinary system and sex hormones, with 4,085 medication-CA combinations recorded for 27% of foetuses. Medications acting on the nervous system (23% of foetuses exposed), antiinfectives for systemic use (22%) and respiratory system medications (17%) were the next most commonly prescribed groups of medications.

Table 4.7. Number of exposures to 523 first trimester medications monitored for signal detection analyses in foetuses with non-chromosomal congenital anomalies (n=15,058) across common ATC2 groups.

Medication group	ATC1 or ATC2 group	Total number of exposures [a]	% of foetuses exposed
Alimentary tract and metabolism	**A**	**2175**	**14.4**
Antacids and medications for peptic ulcer	A02	626	4.2
Medications for functional gastrointestinal disorders	A03	735	4.9
Blood and blood forming organs	**B**	615	4.1
Antithrombotic agents	B01	571	3.8
Cardiovascular system	**C**	**914**	**6.1**
Antihypertensive	C02	215	1.4
Vasoprotective	C05	159	1.1
Beta blocking agents	C07	248	1.6
Calcium channel blockers	C08	185	1.2
Dermatological	**D**	**21**	**0.1**
Genitourinary system and sex hormones	**G**	**4085**	**27.1**
Other gynaecological	G02	943	6.3
Sex hormones	G03	2870	19.1
Systemic hormonal prep., excl. sex hormones and insulins	**H**	**1449**	**9.6**
Thyroid therapy	H03	1157	7.7
Antiinfectives for systemic use	**J**	**3328**	**22.1**
Antibacterial for systemic use	J01	3049	20.2
Antineoplastic and immunomodulating agents	**L**	**115**	**0.8**
Musculoskeletal system	**M**	**558**	**3.7**
Antiinflammatory and antirheumatic products	M01	531	3.5
Nervous system	**N**	**3414**	**22.7**
Analgesics	N02	1730	11.5
Antiepileptics	N03	538	3.6
Psycholeptics	N05	632	4.2
Psychoanaleptics	N06	660	4.4
Antiparasitic products, insecticides and repellents	**P**	**166**	**1.1**
Respiratory system	**R**	**2570**	**17.1**
Nasal preparations	R01	377	2.5
Anti-asthmatics	R03	1253	8.3
Cough and cold preparations	R05	202	1.3
Antihistamines	R06	944	6.3
Various	**V**	**184**	**1.2**
All other therapeutic products	V03	169	1.1

[a] ATC medications with less than 3 exposures across the dataset are not included in counts for this table

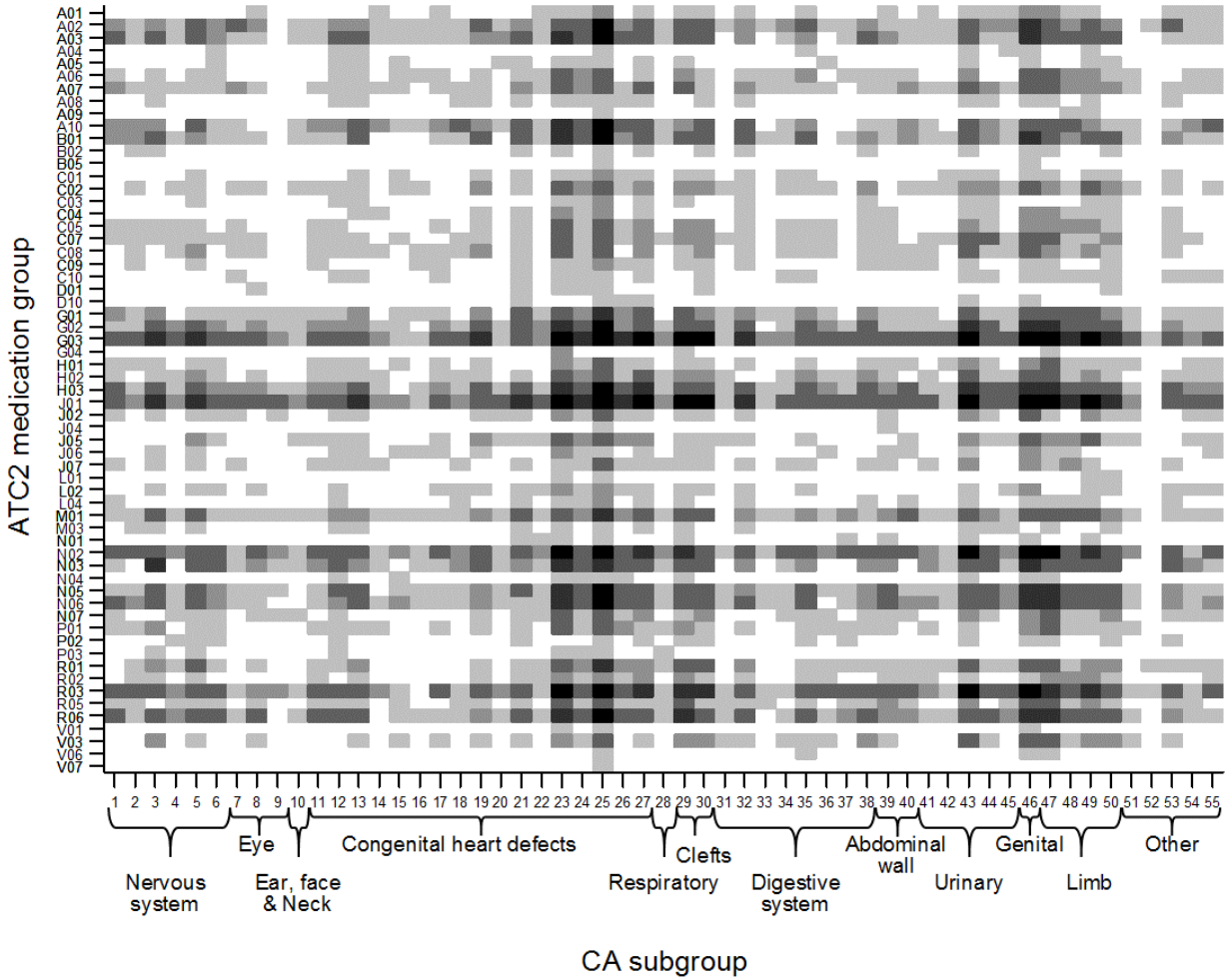
Exposure counts in the crossing of 55 congenital anomalies and 523 medications

When considering the crossing of all 55 CAs and 523 medications considered for signal detection analyses, there were 26,765 exposure counts across 28,765 possible medication-CA combinations. The distribution of these counts is summarised in Table 4.8, where it can be seen that that the majority of combination counts are zero. This means that 77% of the possible combinations between a medication and a CA did not occur in the data. Only 29% of the non-zero medication-CA combinations had three or more exposures recorded. The largest count in the data was 213 exposures for the combination of the most common CHD ventricular septal defect with the thyroid hormone medication H03AA01.

Table 4.8. Distribution of recorded exposure counts in the crossing of 523 medications with 55 congenital anomaly (CA) subgroups.

Exposure count	Number of medication-CA combinations	% of all combinations	% of non-zero combinations
0	22,204	77.2	-
1	3,560	12.4	54.3
2	1,121	3.9	17.1
3	553	1.9	8.4
4	329	1.1	5.0
5-9	591	2.1	9.0
10-49	375	1.3	5.7
50-99	23	0.1	0.4
100-149	5	0.02	0.1
150+	4	0.01	0.1
Total	*28,765*	*100*	*100*

Exposures for 3,355 combinations of 55 CAs with the 61 ATC2 codes are displayed as a heat map in Figure 4.1, where ATC2 groups are used as there is insufficient space to graph the data separately for either the 116 ATC3 groups or the 523 medications. An exposure is a record of one of the 523 ATC5 coded medications taken in combination with one of the 55 CAs. A woman may be counted more than once in either dimension in this data; she may have taken more than one medication (i.e. contributing to more than one row) and/or have had a malformed foetus with more than one CA recorded (i.e. contributing to more than one column). Darker shading in Figure 4.1 shows a more commonly occurring medication-CA combination. Represented by white squares in Figure 4.1, 47% (n=1,582) of all possible combinations of a CA and an ATC2 group of medications had no recorded exposures. Just over half (52%, n=1,735) of the CA and ATC2 combinations had between one and 100 exposures. Only five ATC2-CA combinations had more than 300 exposures, including the ATC2 groups G03 (Sex hormones) and J01 (Antibacterials for systemic use) in combination with atrial and ventricular septal defects, as well as G03 in combination with hypospadias.



Number of Exposures

- 1-5
- 6-10
- 11-50
- 51-100
- >100

CAs with >100 exposures for 1 or more ATC2 group*

Index	Anomaly subgroup	Largest number of exposures in an ATC2 group
25	Ventricular septal defect	636
23	Atrial septal defect	445
46	Hypospadias	356
47	Club foot	204
29	Cleft lip ± palate	186
43	Congenital hydronephrosis	179
49	Polydactyly	144
30	Cleft palate	112
27	Congenital heart defect, unspecified	111

**See Table 4.4 for a full list of index numbers for all 55 CA subgroups*

Figure 4.1. Heat Map of exposure counts for 55 congenital anomalies (CAs) monitored for signal detection, according to ATC2 medication groupings.

Medication exposures per congenital anomaly and congenital anomalies per medication

The number of exposures for each of the 523 medications included in the analysis are displayed in Figure 4.2. These make up the marginal row totals in the two-dimensional crossing of all medications and CAs. The majority of medications (n=471; 90%) had less than 100 counts in combination with any CA. Only seven medications had more than 500 cases of a CA recorded across the data, these were: G03DA04 progesterone (n=1,104), H03AA01 levothyroxine sodium (n=1,017), N02BE01 paracetamol (n=950), G03DB01 dydrogesterone (n=871), J01CA04 amoxicillin (n=787), R03AC02 salbutamol (n=772) and G02CA sympathomimetics (n=556).

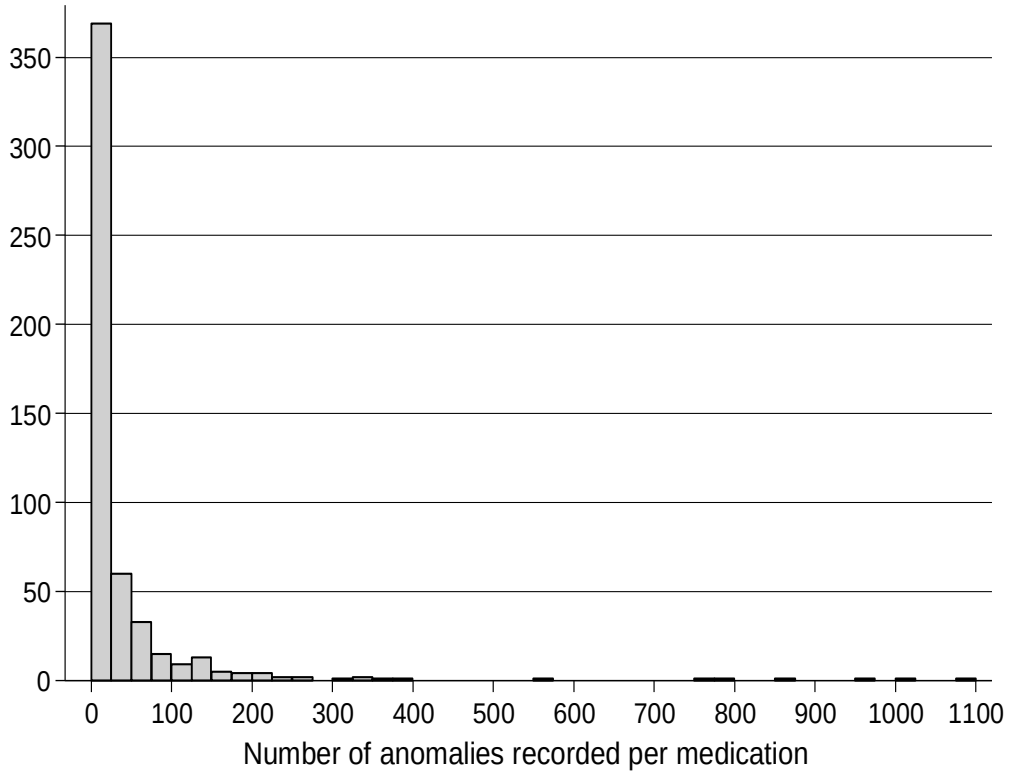


Figure 4.2. Distribution of the number of congenital anomaly subgroups recorded per ATC medication code for 523 medications monitored for signal detection analyses.

In the other dimension, the number of exposures for each of the 55 CAs included in the analysis are displayed in Figure 4.3. These make up the marginal column totals in the two-dimensional crossing of all medications and CAs. The highest number of medication exposures recorded were for ventricular septal defect (n=3,917), atrial septal defect (n=2,033), hypospadias (n=2,022), clubfoot (n=1,297), congenital hydronephrosis (n=1,195) and cleft lip with or without palate (n=1,039), whilst the subgroup conjoined twins had the

least medication exposures (n=6). Over half of the 55 CAs monitored (n=29) had less than 200 recorded exposures to medications monitored for these analyses.

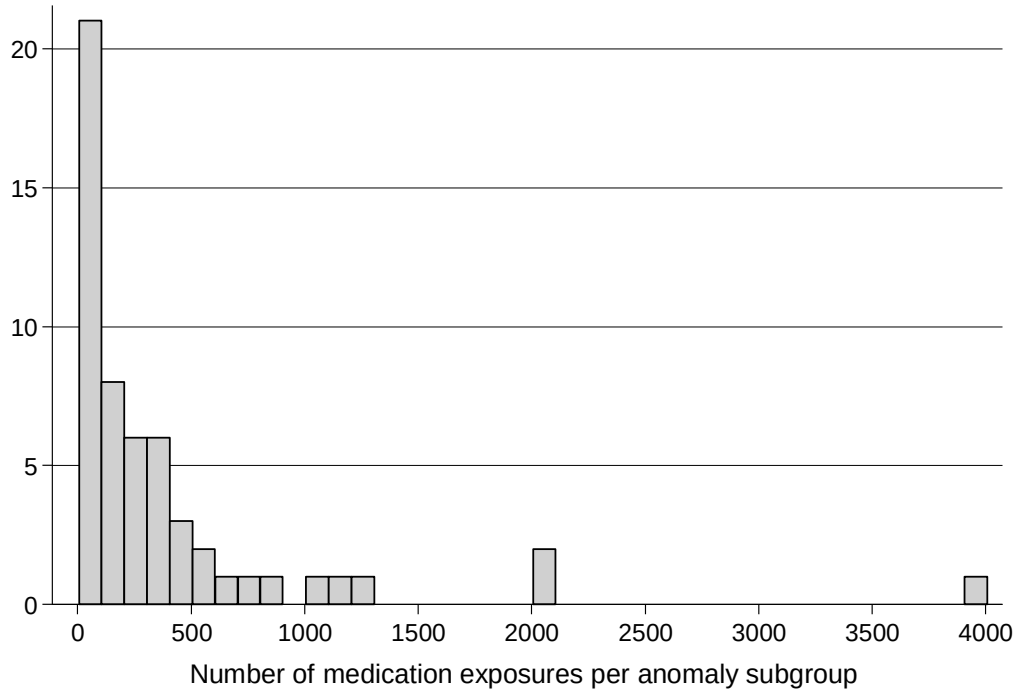


Figure 4.3. Distribution of medication exposures per CA for 55 congenital anomaly subgroups monitored for signal detection analyses.

Foetuses exposed only to medications or CAs not monitored for signal detection analyses

Of the 15,058 women in the dataset, there were 278 and 3,998 who only had medication exposures or CAs, respectively, *other than* the 523 medications or 55 CAs monitored in this thesis. For example, this may include a woman with only one exposure that was a medication for which there were no other exposures recorded in the dataset, because such a medication would not be monitored for signal detection analyses with an overall count in the data of less than three. Another example would be a malformed foetus with only one recorded CA, where that CA was included in the dataset but not monitored separately for signal detection analyses, e.g. cystic adenomatous malformation of lung. These individuals were included in the counts as "controls" and hence they contributed to the marginal total counts in the model.

4.4.3. Merging information from the Australian risk categorisation database with EUROmediCAT data

Table 4.9 displays the breakdown of the risk categories for the 523 medications and the 28,765 medication-CA combinations for ATC4 and ATC5 codes in the EUROmediCAT dataset. The majority of ATC5 codes were identified as belonging to one of the risk categories. Very few medications coded only to ATC4 level were specified in the Australian categorisation database; however, ATC4 codes were only considered in the analysis for cases where a more specific ATC5 code was not available (less than 5% of exposures). Three medications (chloroquine, cyproterone and medroxyprogesterone) could be mapped to a code in both "low risk" and "high risk" groups depending on the dosage at which they were used. As there is no information on dosage in EUROmediCAT data, these medications were not assigned to either risk category, and were instead coded as "no risk category identified". Overall, 65% of the 28,765 combinations available for signal detection analyses were mapped to one of the risk categories. Only 44 (8.4%) of the medications (or the medication-CA combinations) were assigned to the "high risk" category. It is important to note that the number of medication-CA combinations identified as "high risk" may be overestimated since these categorisations are specific only to the medications and not the type of CA. This is likely to be the case as most medications will carry a higher risk of only a few specific CAs, rather than an increased risk for all CAs in general. The number of combinations in this data that will be true signals of teratogens is therefore likely to be considerably less than 8.4%.

Table 4.9. Number of medication-congenital anomaly (CA) combinations in the "Low" and "high" risk categories, for 28,765 potential medication-CA combinations.

Risk category	**ATC4-CA combinations**		**ATC5-CA combinations**		**Total combinations**		**Total medications**	
	N	%	N	%	N	%	N	%
No risk category identified	2,090	97.5	7,920	29.7	10,010	34.8	182	34.8
Low risk: Category A, B or C	55	2.5	16,280	61.2	16,335	56.8	297	56.8
High risk: Category D or X	0	0	2,420	9.1	2,420	8.4	44	8.4
Total	*2,145*		*26,620*		*28,765*		*523*	

4.5. Summary

This chapter has discussed signal detection methods used in the context of SR databases and in the context of first trimester medications and CAs. In the following two chapters the identified approaches to try and improve current EUROmediCAT signal detection methods are investigated and assessed. This was done by considering groupings of medications and/or CAs, using two approaches to signal detection analyses that are new for CA data. Chapter 5 presents different FDR procedures that group similar medications and/or CAs when determining the statistical significance of each test, and chapter 6 investigates the grouping of similar medications and/or CAs using BHMs.

Chapter 5: Analysis of EUROmediCAT safety of medication use during pregnancy I: false discovery rate

5.1. Introduction

This chapter begins by describing a number of approaches to FDR multiple testing procedures that consider groupings of medication-CA combinations when determining the statistical significance of each test. Results from these analyses are directly compared to those from an FDR procedure that does not consider any such groupings, as this is the current approach used by EUROmediCAT. In order to directly compare the methods proposed here with those used currently, methodology from the recent EUROmediCAT signal detection analyses [Luteijn et al., 2016] was used. The main aim of these approaches was to determine whether the signal detection process recently developed by EUROmediCAT could be improved at a basic level by considering an alternative multiple testing procedure when adjusting P-values post-analysis.

5.2. Methods

A one-sided Fisher's exact test of association was performed separately for each medication-CA combination in this "exposed-malformed" study design (see Table 4.1). This was followed by one of four FDR procedures (described in the following section) to adjust for multiple testing when determining statistical significance of each combination. Since the dataset consisted only of CA cases, the power to detect protective associations was very low. Furthermore, since the purpose of signal detection analyses is to screen for potentially harmful teratogenic medicines, any preventive associations that are identified by these methods would not be flagged as signals or be recommended for further examination. Therefore, a one-sided hypothesis test was used for each medication-CA combination rather than a two-sided test. Only medications with at least 3 exposures were investigated, although foetuses exposed only to medications not reaching this threshold were included in the data as controls. Furthermore, whilst combinations of medications and CAs with less than three exposed cases were included in the multiple testing, these were not considered in the resulting set of potential signals. A lack of low powered associations would violate the underlying assumptions of the multiple testing procedure, as this markedly shifts the distribution of P-values towards zero, hence these associations were retained in the multiple testing procedures but were not flagged as signals for further consideration.

5.2.1. False discovery rate procedures

Four variants on the FDR procedure were considered in this chapter; firstly, a Simes procedure [Benjamini and Hochberg, 1995] that did not consider any groupings was applied, as this is the existing methodology used in EUROmediCAT signal detection. Three FDR methods that incorporate groupings of medications or CAs were then considered. The way in which these four FDR procedures were implemented is described below.

*Benjamini-Hochberg false discovery rate procedure (**single FDR**)*

For a total of m tests and tolerating that a given proportion α of discoveries will be false, the basic Benjamini-Hochberg, Simes or **single FDR** procedure [Benjamini and Hochberg, 1995], is specified as follows

- Order the P-values P_i according to their magnitude $P_1 < \cdots < P_m$
- Let $\tilde{P}_i$ denote the corresponding FDR adjusted value for P_i, where

$$\tilde{P}_m = P_m$$

$$\tilde{P}_i = min\left(\frac{m}{i}P_i\,, \tilde{P}_{i+1}\right) \quad for\ i\ \leq m-1$$

- Null hypotheses with $\tilde{P}_i \leq \alpha$ are rejected; these combinations are the potential signals

The FDR "cut-off" is defined for a pre-specified proportion α between 0 and 1; for example, $\alpha = 0.1$ corresponds to an FDR of 10%, which means that up to 10% of combinations that are identified as signals are expected to be false positives. FDR-adjusted P-values are calculated in ascending order of magnitude; if an FDR-adjusted P-value is 0.1 this means that 10% of all the combinations with an adjusted P-value less than this might be false positives. Benjamini and Hochberg [1995] showed that for m independent tests, this controls the FDR at the level $\frac{m_0}{m}\alpha \leq \alpha$, where m_0 is the number of true null hypotheses.

*False discovery rate procedure by groups (**FDR by group**)*

The **FDR by group** method uses a two-step procedure in which a number of groups are first eliminated from the set of P-values, and then each remaining group of medications is considered separately, with P-values adjusted within each group. This method will increase the number of signals identified compared to single FDR, but also increases the overall proportion of these signals that are likely to be false positive associations. In order to reduce the dimension of the data, a first step in this process is to disregard groups of medications that do not have any associations below a certain level of significance. The

procedure is based on the original "double FDR" procedure of Mehrotra and Heyse [2004], which groups similar AEs together in the safety analysis of clinical trials data. The original double FDR procedure is implemented only after an initial step, in which all of the rare AEs are eliminated from the data. This is done for the purposes of dimension-reduction as well as to exclude those rare AEs that could not possibly reach statistical significance even at the conventional 5% significance level without any adjustment for multiple testing. In the original double FDR procedure, the smallest P-value within each group is first identified and then a single FDR is performed across this set of representative minimum P-values from each group; only groups with an FDR-adjusted minimum P-value below a certain significance level α_1 are considered further. In a second stage, FDR adjustments are then done within each group separately, with a significance level α_2. However, this method requires nonparametric bootstrapping to determine the optimal values for the cut-offs α_1 and α_2 in order to minimise the overall FDR level α for the particular scenario under consideration. Mehrotra and Heyse determined that a reasonable choice for controlling the overall FDR was to choose values of $\alpha_1 = \frac{\alpha_2}{2}$, which they demonstrated maintained an overall FDR at level $\alpha = \alpha_2$. This was suggested as an ad-hoc alternative to bypass the bootstrap resampling [Mehrotra and Heyse, 2004]. For this chapter, therefore, $\alpha_1$was set to $\frac{\alpha_2}{2}$, for various levels of α_2. In practice, the groups that are thrown out in the first stage of this process would be unlikely to include any potential signals after FDR adjustment to P-values in the second stage, even at the increased cut-off level of $\alpha_2 = 2\alpha_1$. This procedure therefore provides similar results as would be obtained from a separate FDR procedure for each group (i.e. not excluding any groups first, as is done here).

The **FDR by group** procedure was specified for $i = 1, \dots, n$ groups each with a total of g_i $(j = 1, \dots, g_i)$ medications as follows

Step 1: Let P_i be the minimum P-value from each group of medications i

$$P_i = \min(P_{ij}; 1 \leq j \leq g_i)$$

and order the P_i by magnitude for $i = 1, \dots, n$ where $P_1 < \cdots < P_n$.

Then $\tilde{P}_i$ denote the corresponding FDR adjusted P-values, where

$$\tilde{P}_n = P_n$$

$$\tilde{P}_i = min\left(\frac{n}{i}P_i, \tilde{P}_{i+1}\right) \quad for\ i \leq n-1$$

Step 2: Apply further FDR adjustments to the g_i P-values within each group i for which $\tilde{P}_i \leq \alpha_1$, such that $\tilde{P}_{ij}^{(i)}$denote the FDR-adjusted P-values for each P-value

in i. Then null hypotheses are rejected for all tests where $\tilde{P}_{ij}^{(i)} \leq \alpha_2$; these combinations are the potential signals.

*Group Benjamini-Hochberg method (**Group BH**)*

Another FDR method that enables groupings to be considered is the group Benjamini-Hochberg (group BH) method of Hu et al. [2010]. Intended for use in large-scale applications such as genome-wide association studies, this is a two-step method that weights P-values in a first step and then applies an FDR adjustment to these weighted P-values in a second step. The idea behind this approach is that those groups in which there are proportionally more true signals are given a greater statistical weight in the adjustment to P-values. The actual proportion of true signals is, of course, unknown, and must therefore be estimated.

The two-step **group BH** procedure can be summarised for $i = 1, \ldots, n$ groups each with a total of g_i ($j = 1, \ldots, g_i$) medications as follows

Step 1: For the P-values within each group i, apply a single FDR procedure at level $\alpha' = \frac{\alpha}{1+\alpha}$ and estimate the true number of null hypotheses π_{0i} (i.e. for which there is no true association) in group i using the number of null hypotheses φ_{0i} in group i that are rejected at level α'. Then

$$\hat{\pi}_{0i} = \frac{\varphi_{0i}}{g_i}, \qquad \hat{\pi}_0 = \frac{\sum_{i=1}^{n} \varphi_{0i}}{N}, \quad N = \sum_{i=1}^{n} g_i$$

Then the weighted P-value for combination ij is defined as

$$P_{ij}^{w} = \frac{\hat{\pi}_{0i}}{1 - \hat{\pi}_{0i}} P_{ij}, \qquad \text{where } P_{ij}^{w} = \infty \ if \ \hat{\pi}_{0i} = 1$$

Step 2: The weighted P-values for all groups with $\hat{\pi}_{0i} < 1$ are pooled together and a second FDR adjustment is performed across the remaining set of weighted P-values to get $\tilde{P}_{ij}^{w}$, the FDR-adjusted value corresponding to P_{ij}^{w}. A signal is then flagged for all $\tilde{P}_{ij}^{w} \leq \alpha^{w}$, where

$$\alpha^{w} = \frac{\alpha}{(1+\alpha)(1-\hat{\pi}_0)}$$

Note that in the extreme case that the proportion $\hat{\pi}_{0i} = 1$, all P-values in that group are re-scaled to infinity to ensure that no combinations are flagged as signals in a group where it is estimated that all null hypotheses are true, and the attention is therefore focused on groups for which $\hat{\pi}_{0i} < 1$. If $\hat{\pi}_{0i} = 1$ for all i then no null hypotheses are rejected.

*Double false discovery rate procedure (**Double FDR**)*

In a more recent paper, Mehrotra and Adewale [2012] set out an updated double FDR procedure, in which implementation of their original double FDR method was simplified by eliminating the need to perform bootstrap sampling to find the best choice of α_1 and α_2. Demonstrating better control of the overall FDR, the key difference compared to the original double FDR procedure [Mehrotra and Heyse, 2004] is in the choice for the representative P-values P_i for each group in the first step. Various choices for the representative P-value P_i^* for each group i are discussed, with the recommend choice for P_i^* (see below) being shown to provide the highest power to detect signals whilst maintaining a similar FDR to other options considered [Mehrotra and Adewale, 2012].

This method is referred to as the **double FDR**, and is specified for $i = 1, \dots, n$ groups each with g_i $(j = 1, \dots, g_i)$ medications as follows

Step 1: First perform a single FDR adjustment within each group i to get $\tilde{P}_{ij}$, then let P_i^* denote the smallest FDR-adjusted P-value in each group

$$P_i^* = \min(\tilde{P}_{ij}; 1 \le j \le g_i)$$

Then apply an FDR adjustment to the P_i^* to get the set of representative FDR-adjusted P-values $\tilde{P}_i^*$, where $P_1^* < \cdots < P_n^*$ and

$$\tilde{P}_n^* = P_n^*$$

$$\tilde{P}_i^* = min\left(\frac{n}{i} P_i^* , \tilde{P}_{i+1}^*\right) \quad for\ i \le n-1$$

Then all groups i for which $\tilde{P}_i^* \le \alpha$ are taken to the second step of the procedure. P-values from groups with $\tilde{P}_i^* > \alpha$ are not considered further.

Step 2: Let $F \equiv \{P_{ij} \mid \tilde{P}_i^* \le \alpha\}$ be the family of all P-values from groups flagged by their representative FDR-adjusted P-values $\tilde{P}_i^*$ in step 1. Then apply a single FDR procedure across all P-values in F such that $\tilde{P}_{ij}^{(F)}$ is the FDR-adjusted P-value for all $P_{ij} \in F$. Then null hypotheses are rejected for all tests where $\tilde{P}_{ij}^{(F)} \le \alpha$; these combinations are the potential signals.

Note that the key difference between the original double FDR of Mehrotra and Heyse (here the FDR by group procedure) and the "new" double FDR is that the adjustment in the second step of the newer procedure is made across *all* P-values in the groups remaining at the same time, rather than separately within each remaining group. The actual overall FDR level for this procedure is at most α at the group level, no matter how many groups contain

at least one true signal, hence it may be possible that in practice the overall FDR exceeds α in some scenarios [Mehrotra and Adewale, 2012].

5.2.2. Groupings used for false discovery rate methods

Medications

Groupings of medications were defined using the ATC coding hierarchy. Firstly, this was done according to the ATC2 level, which classifies drugs according to the main therapeutic use of their main active ingredient. One such group, for example, would be A10 "Drugs used in diabetes", within which are nested 85 specific ATC5 codes. The next level of grouping medications was to use the ATC3 level classification of pharmacological subgroup. In the previous example, instead of having the 85 medications that fall under the ATC2 classification A10 together as one group, for example, there were 3 separate groupings as follows: A10A "insulins and analogues", A10B "blood glucose lowering drugs, excluding insulins" and A10X "other drugs used in diabetes". Grouping using ATC4 level classification of chemical subgroup was also considered as a sensitivity analysis. This choice of grouping was not included in the main results as it was expected to provide too many groups to be useful for these analyses (since there would be many groups and each with only a small number of medications within them). The only other ATC level that could potentially have been used for grouping was ATC1, however this describes the 14 anatomical main groups and these were too broad to be useful here as each resulting group would include such a large number and variety of different medications.

Congenital anomalies

Another way of considering groupings in the adjustment for multiple testing was to use the 55 CAs to group medication-CA combinations, such that one group would then include the combination of each of the 523 medications with a single CA. This could make use of the fact that there are likely to be more signals within certain CAs, as well as taking account for the fact that a number of subgroups may have had no signals at all (i.e. contain only null hypotheses that would not be rejected).

5.2.3. Summary of methods used in this chapter

In summary, groupings by ATC2 and ATC3 medication codes and by CA were considered in the application of four approaches to adjustment for multiple testing to determine the statistical significance of each test:

1. FDR across all tests (**single FDR**)
2. FDR procedure separately within each group, after excluding groups in a first step (**FDR by group**)
3. Two-step weighted FDR procedure considering groupings (**group BH**)
4. Double FDR procedure considering groupings (**double FDR**)

The PRR for each medication-CA combination was plotted against the Fisher's exact test P-value using a smileplot [Newson, 2003]. As described in chapter 4, results from all models were compared in terms of their identification rate and "high risk" proportion according to the Australian risk categorisation system for ATC coded medications. Briefly, these measures were defined as

$$\textbf{Identification rate} = \frac{\text{Number of "high risk" medications identified as signals}}{\text{Total number of "high risk" medications in the data}}$$

$$\text{"}\textbf{High risk}\text{" }\textbf{proportion} = \frac{\text{Number of medication signals in "high risk" category}}{\text{Total number of medications identified as signals}}$$

$$\textbf{Effective workload} = \text{Total number of medication signals}$$

5.3. Results

For a description and summary of the EUROmediCAT dataset used for these analyses see Chapter 4 (section 4.4.2). Briefly, data on 15,058 foetuses was available for analysis, with 523 ATC medications and 55 CAs being monitored for signal detection purposes giving rise to 26,765 total exposure counts across 28,765 possible medication-CA combinations.

5.3.1. Fisher's exact test and a single FDR procedure

There was no data to perform an analysis for 369 medication-CA combinations for which there were no registries with at least one record of both that specific medication exposure and the specific CA. Fisher's exact test was therefore performed for 28,396 medication-CA combinations, the results of which are displayed in a smileplot in Figure 5.1. Each marker on this plot represents a different medication-CA combination, with different symbols used to show the risk category of the medication in each combination. The dashed horizontal lines show some of the possible cut-off values of α for the FDR, and the solid vertical line indicates a PRR of one (i.e. no effect). The points above each dashed line and to the right of the vertical line correspond to those medication-CA combinations that would be flagged as potential signals using that particular FDR cut-off. Using single FDR, only one "high risk" combination out of 2 total combinations were identified as signals for an FDR cut-off of 5% (points above the uppermost dashed line in Figure 5.1). For FDR cut-offs of 20% and 50% there was one "high risk" out of 4 total signals and 5 "high risk" out of 10 total signals, respectively. This means that only 5 (11%) of the 44 "high risk" medications in the data were identified as a signal when using single FDR with the highest cut-off of 50%. Note that the count for the FDR 50% does not include the two points above the dashed FDR 50% line that are to the left of the vertical solid line, since these estimates have a PRR of less than one, indicating a "protective" association for that particular medication-CA combination. If the data were analysed without any consideration for multiple testing and at a significance level of 5% (i.e. using the P-value cut off of 0.05) then n=252 combinations would be identified as signals, after excluding those combinations with a frequency of less than 3 or a PRR below 1 (n=391 and n=27, respectively).

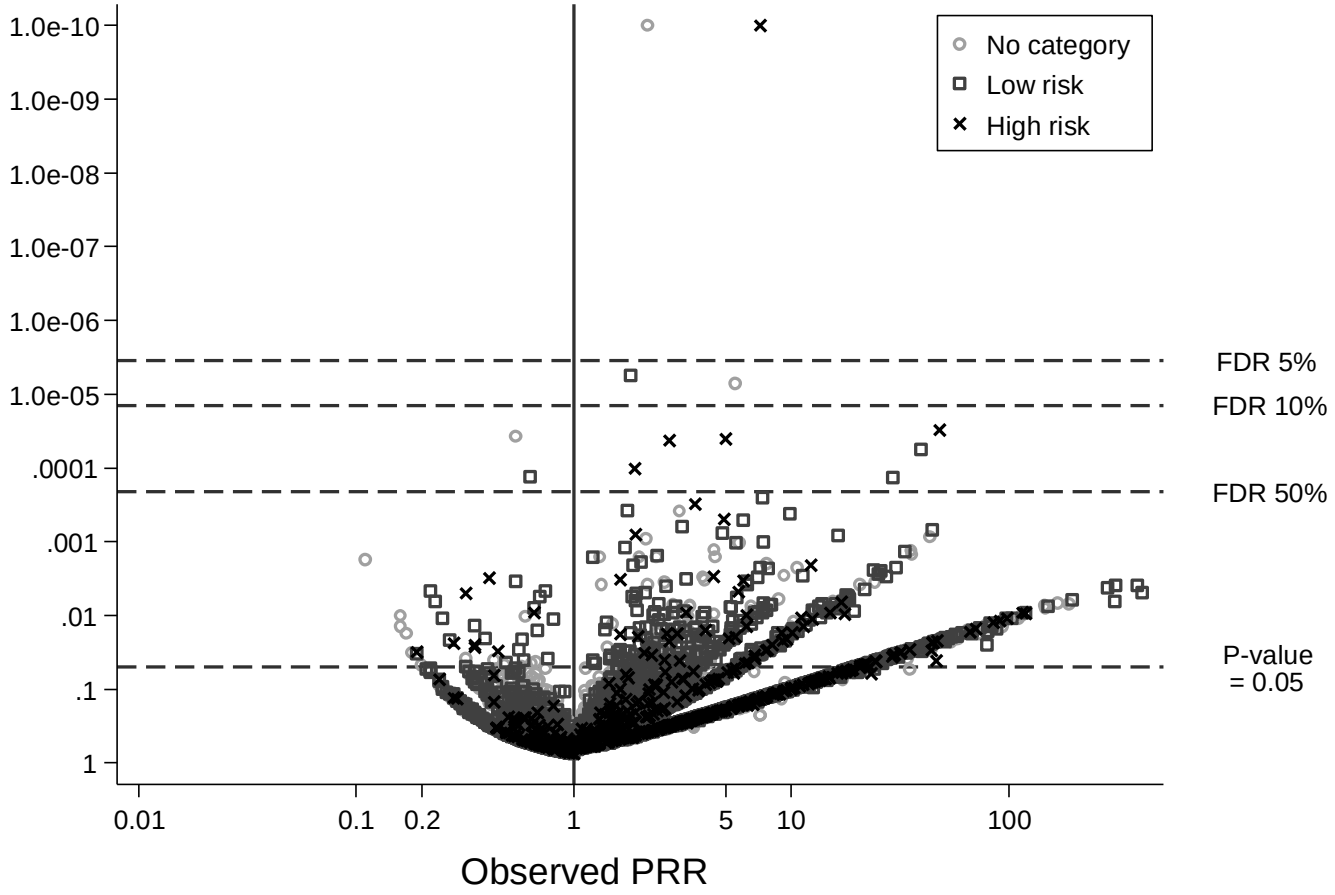


Figure 5.1. Smile plot of the observed PRR against the unadjusted P-value from Fisher's exact test for 28,396 medication-CA combinations, with different shading/symbols according to Australian risk categories. Single FDR cut-off levels are indicated by dashed horizontal lines and two P-values of 1.4e-17 and 2.0e-17 are shown at P=1.0e-10 for illustration purposes.

5.3.2. False discovery rate procedures considering groupings of medications or congenital anomaly subgroups

Before comparing the FDR procedures directly, the way in which combinations could potentially be grouped was first considered for all 3 methods that took grouping into account. Grouping by ATC2 provided 61 groups with an average of 9 (range 1-54) medication codes and 1,074 (range 53-2,936) medication-CA combinations per group. Using ATC3 grouping resulted in 116 distinct groups with an average of 4.5 (range 1-20) unique medications and 487 (range 53-1,086) medication-CA combinations per group. Grouping by CA gave 55 distinct groups with an average of 518 (range 325-523) unique medication-CA combinations per group. Note that the number of combinations was not a multiple of 55 (for medication groupings) or 523 (for grouping by CA) in all of the groups due to the 369 combinations for which there was no data to perform Fisher's exact test after excluding registries without that specific exposure or CA.

Number of signals identified using ATC2, ATC3 and CA groupings

Table 5.1 displays the number of signals of unique ATC coded medications detected using each type of FDR procedure and grouping, with separate rows for each method over a range of cut-off levels from 5% to 50%. All counts in Table 5.1 exclude those associations found to be significant (after the relevant FDR procedure) that had less than 3 exposures for that particular medication-CA combination and those with a PRR<1 (i.e. protective associations). The number of groups without any signals (i.e. in which no null hypotheses were rejected) after each FDR procedure is shown in the first set of columns; on average, around 80-90% of groups contained no signals after the FDR procedures. The lowest proportion of groups with no signals was seen for the highest FDR cut-off of 50%, in particular for CA groupings using FDR by group, where only 42 (76%) of the CA groups did not have any potential signals. The highest proportion was seen for the lowest FDR cut-offs of 5% (single and double FDR) and 10% (double FDR) where 114 (98%) of ATC3 groups had no signals. The second set of columns in Table 5.1 show the number of medication signals that are in the "high risk" category and the final set of columns gives the total number of medications identified as potential signals. Note that the number of signals is the same for all three columns for single FDR, as this method does not consider any groupings. For all methods and groupings, the number of "high risk" and the total number of signals both increased with the FDR cut-off.

Table 5.1. Summary of the number of medications identified as signals (effective workload) for all FDR methods and groupings.

FDR cut-off	FDR method	Groups with no signals after FDR			"High risk" medication signals			Total number of medication signals		
Type of grouping (number of groups)		*ATC2 (61)*	*ATC3 (116)*	*CA (55)*	*ATC2*	*ATC3*	*CA*	*ATC2*	*ATC3*	*CA*
5%	Single FDR	59	114	53	——— 1 [a] ———			——— 2 [a] ———		
	FDR by group	55	110	48	3	3	4	7	7	9
	Group BH	55	110	48	3	3	4	7	7	9
	Double FDR	59	114	53	2	2	3	3	3	5
10%	Single FDR	57	112	51	——— 1 ———			——— 3 ———		
	FDR by group	54	108	46	5	5	5	10	14	12
	Group BH	54	108	46	5	5	5	10	14	12
	Double FDR	59	114	51	2	4	3	3	5	11
20%	Single FDR	55	110	48	——— 3 ———			——— 7 ———		
	FDR by group	53	104	45	5	6	5	14	20	18
	Group BH	53	106	46	5	6	5	14	18	16
	Double FDR	56	112	50	4	5	5	11	11	13
30%	Single FDR	55	110	47	——— 3 ———			——— 7 ———		
	FDR by group	52	101	44	5	7	7	16	26	30
	Group BH	53	104	45	5	6	5	14	20	19
	Double FDR	55	109	48	5	6	5	12	15	18
40%	Single FDR	54	109	46	——— 3 ———			——— 8 ———		
	FDR by group	51	98	43	5	8	7	18	31	40
	Group BH	52	101	44	5	7	7	15	25	29
	Double FDR	55	109	47	6	6	5	16	15	20
50%	Single FDR	54	109	46	——— 3 ———			——— 8 ———		
	FDR by group	48	96	42	7	8	8	28	34	55
	Group BH	53	100	43	5	7	7	16	27	38
	Double FDR	53	109	46	7	6	5	21	16	24

[a] There is no grouping for the single FDR method, hence numbers of signals are the same for all three columns

Effective workload according to risk categories for ATC2, ATC3 and CA groupings

Figure 5.2, Figure 5.3 and Figure 5.4 show the effective workload broken down by risk categories for the FDR by group, group BH and double FDR, respectively, comparing each type of grouping (ATC2, ATC3 or CA) for a range of FDR cut-offs in 5% increments from 5% to 50%. For FDR by group (Figure 5.2), the number of "high risk" medication signals was similar for all three groupings, although ATC3 groupings identified more "high risk" signals for some levels of FDR cut-off. The overall number of signals identified was lowest for ATC2 groupings, and was notably higher for grouping by CA. For group BH (Figure 5.3), ATC3 grouping consistently identified slightly more signals up to an FDR of 40%, past which the CA groupings provided more signals but without any increase in the number of "high risk" signals. The number of "high risk" signals identified by each type of grouping was similar throughout; grouping by ATC2 resulted in slightly fewer "high risk" signals for FDR cut-offs of 20% and higher. For double FDR (Figure 5.4), ATC3 groupings identified a similar number of "high risk" signals but often with less signals overall compared to other groupings. For an FDR cut-off of 50% the ATC2 groupings identified one additional "high risk" medication, however this was for an additional five medication signals in total.

"High risk" proportion and identification rate for ATC2, ATC3 and CA groupings

Figure 5.5 shows the "high risk" proportion plotted against the effective workload for the three groupings and methods. There are ten points for each type of grouping, with each point representing a different level of FDR cut-off in a sequential order by increments of 5% from 5% to 50%. For each method, therefore, the points in Figure 5.5 with the lowest and highest total number of identified signals correspond to the FDR of 5 and 50%, respectively. A higher point on Figure 5.5 indicates an improvement in "high risk" proportion, i.e. that a greater percentage of the set of signals requiring follow up are in the "high risk" category. For FDR by group and group BH the performance of the three groupings was very similar. Grouping by CA performed worse when identifying around 30 or more medication signals for FDR by group and for most levels of double FDR. The double FDR with ATC3 groupings generally had the highest percentage of signals identified being in the "high risk" group for the lowest total number of signals across different levels of FDR, although ATC2 groupings had only marginally inferior performance.

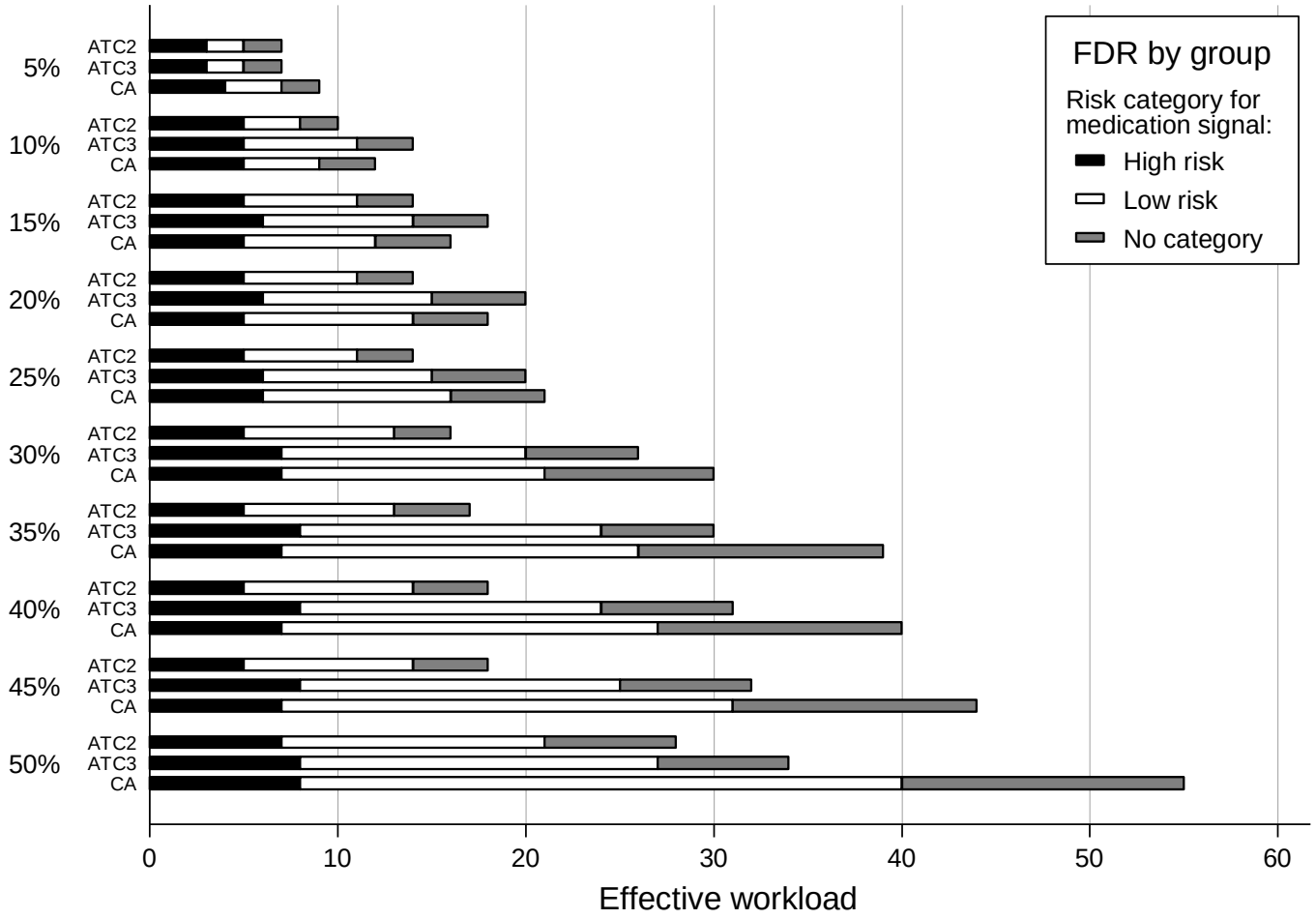


Figure 5.2. Effective workload and the number of medication signals in each risk category using the FDR by group procedure. Results are for grouping of medication-CA combinations by ATC2, ATC3 codes and CAs according to cut-offs for FDR level from 5% to 50%.

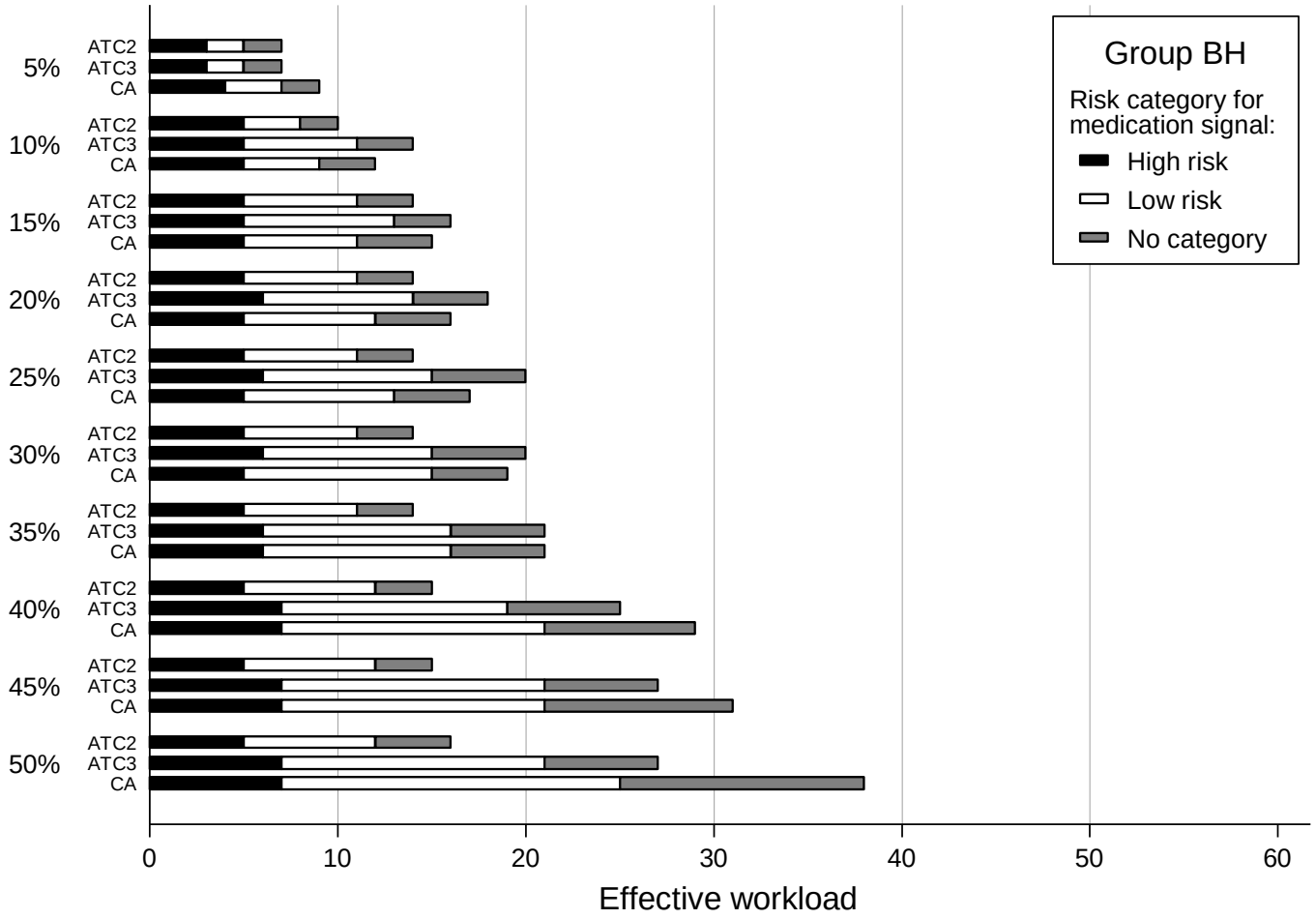


Figure 5.3. Effective workload and the number of medication signals in each risk category using the group BH procedure. Results are for grouping of medication-CA combinations by ATC2, ATC3 codes and CAs according to cut-offs for FDR level from 5% to 50%.

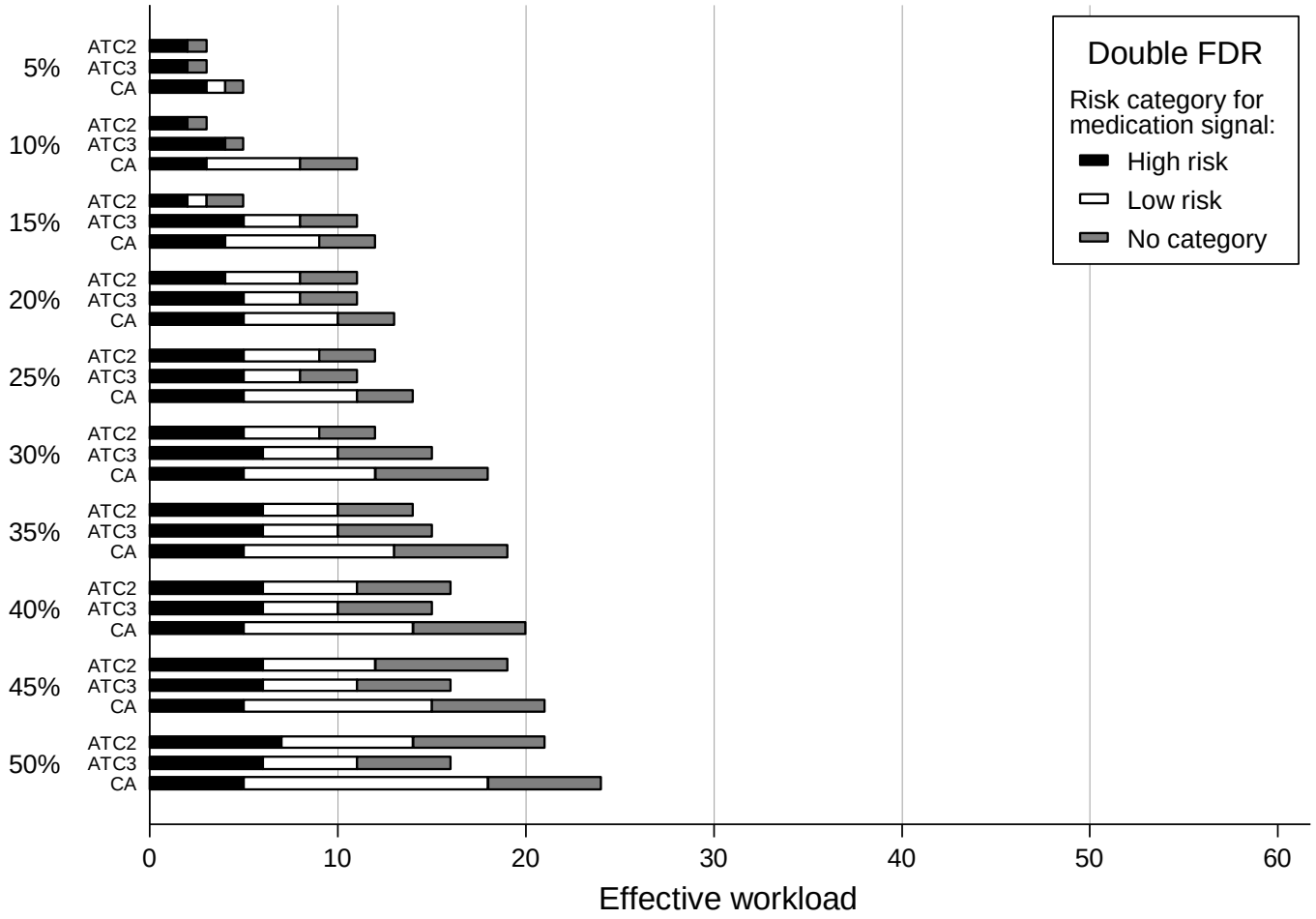


Figure 5.4. Effective workload and the number of medication signals in each risk category using the double FDR procedure. Results are for grouping of medication-CA combinations by ATC2, ATC3 codes and CAs according to cut-offs for FDR level from 5% to 50%.

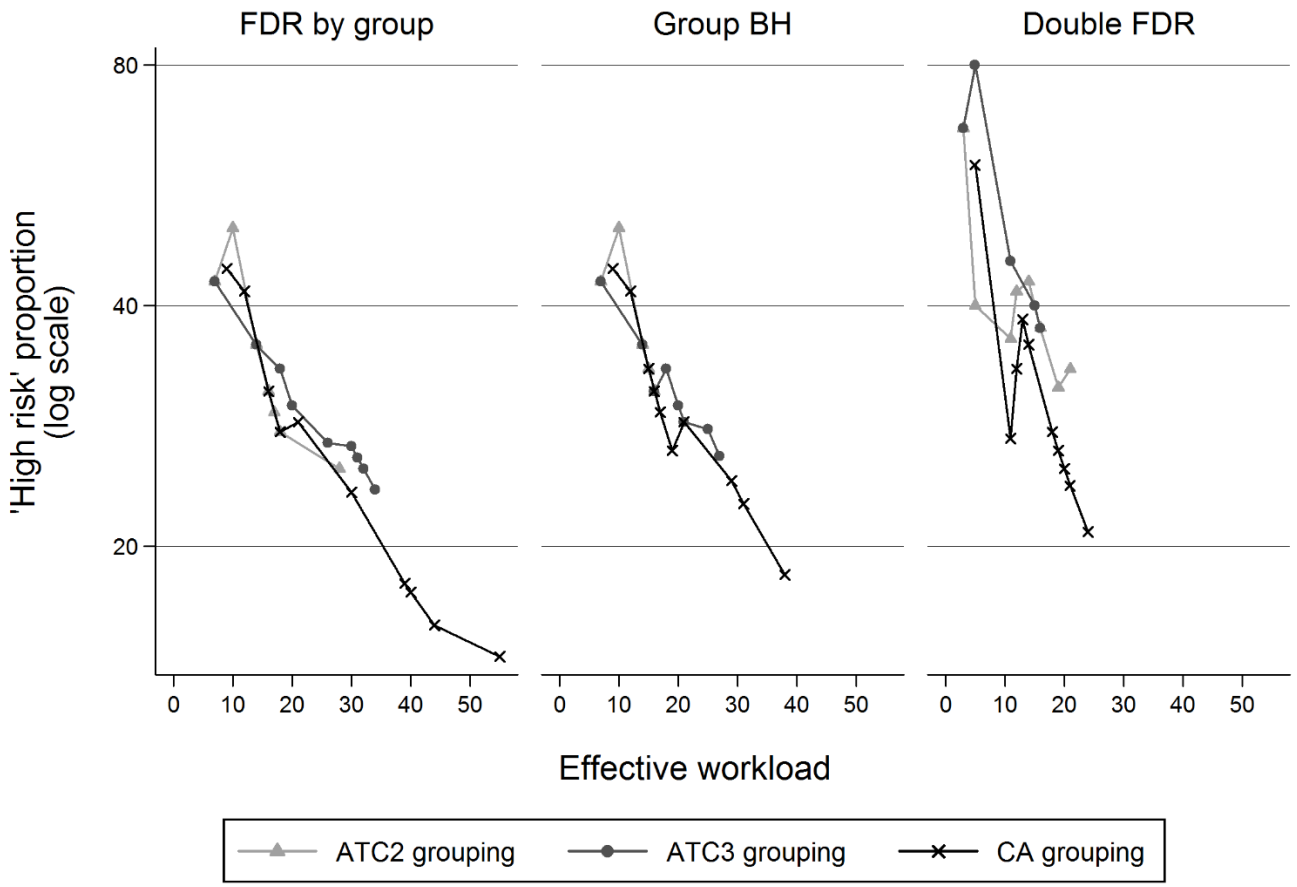


Figure 5.5. "High risk" proportion (percent of all medication signals that are in the "high risk" category) against effective workload (the total number of medication signals) for FDR by group, group BH and double FDR methods. Grouping is by ATC2, ATC3 codes and by CA, and each point corresponds to a different level of FDR for that type of grouping (in 5% increments from 5% to 50%).

The identification rate is plotted against the effective workload in Figure 5.6 for the three groupings and methods, again with ten points for each type of grouping, each representing a level of FDR cut-off from 5% to 50%. In Figure 5.6, points lying closer to the top left hand corner of the graph indicate a higher identification rate for a smaller total number of signals requiring follow up i.e. representing a better signal detection scenario. For FDR by group and group BH all the groupings performed similarly overall. For double FDR, ATC3 grouping had slightly improved performance compared to ATC2 grouping, which in turn had higher identification rates for the same effective workload as grouping by CAs. When using CA groupings, the double FDR showed no improvement in identification rate for FDR cut-offs of above 20%; the effective workload increased without any additional "high risk" medication signals being identified.

Sensitivity analysis: false discovery rate methods using ATC4 groupings

As a sensitivity analysis, FDR procedures were repeated using the ATC4 chemical subgroups as the groupings. Groupings by ATC4 gave 245 groups with an average of 2.1 (range 1-8) unique medication codes and 184 (range 42-440) medication-CA combinations per group. The ATC4 groupings for FDR by group and group BH methods gave a larger number of overall signals and effective workload than grouping by ATC2 or ATC3, generally without any increase in the number of "high risk" medications being detected (Appendix Figure B1 and Figure B2). Only CA groupings at a cut-off level of 45-50% identified more overall signals then ATC4 groupings. Conversely, when using double FDR the ATC4 groupings gave the smallest overall total numbers of signals, with considerably less "high risk" medications being identified as signals compared to the other ATC groupings (Appendix Figure B3). When considering the "high risk" proportion, identification rates and effective workloads, grouping combinations by ATC4 level codes did not show any improvement over other groupings considered (Appendix Figures B4 and Figure B5).

Summary of false discovery rate methods by ATC2, ATC3 and CA groupings

Overall, for the FDR by group and group BH method the three groupings gave similar results. For the double FDR method, however, the ATC3 groupings performed slightly better in terms of "high risk" proportion and identification rate across the range of cut-off choices for the FDR. In the following section, therefore, the focus will be on ATC3 groupings when directly comparing the FDR methods that group medication-CA combinations to the single FDR method.

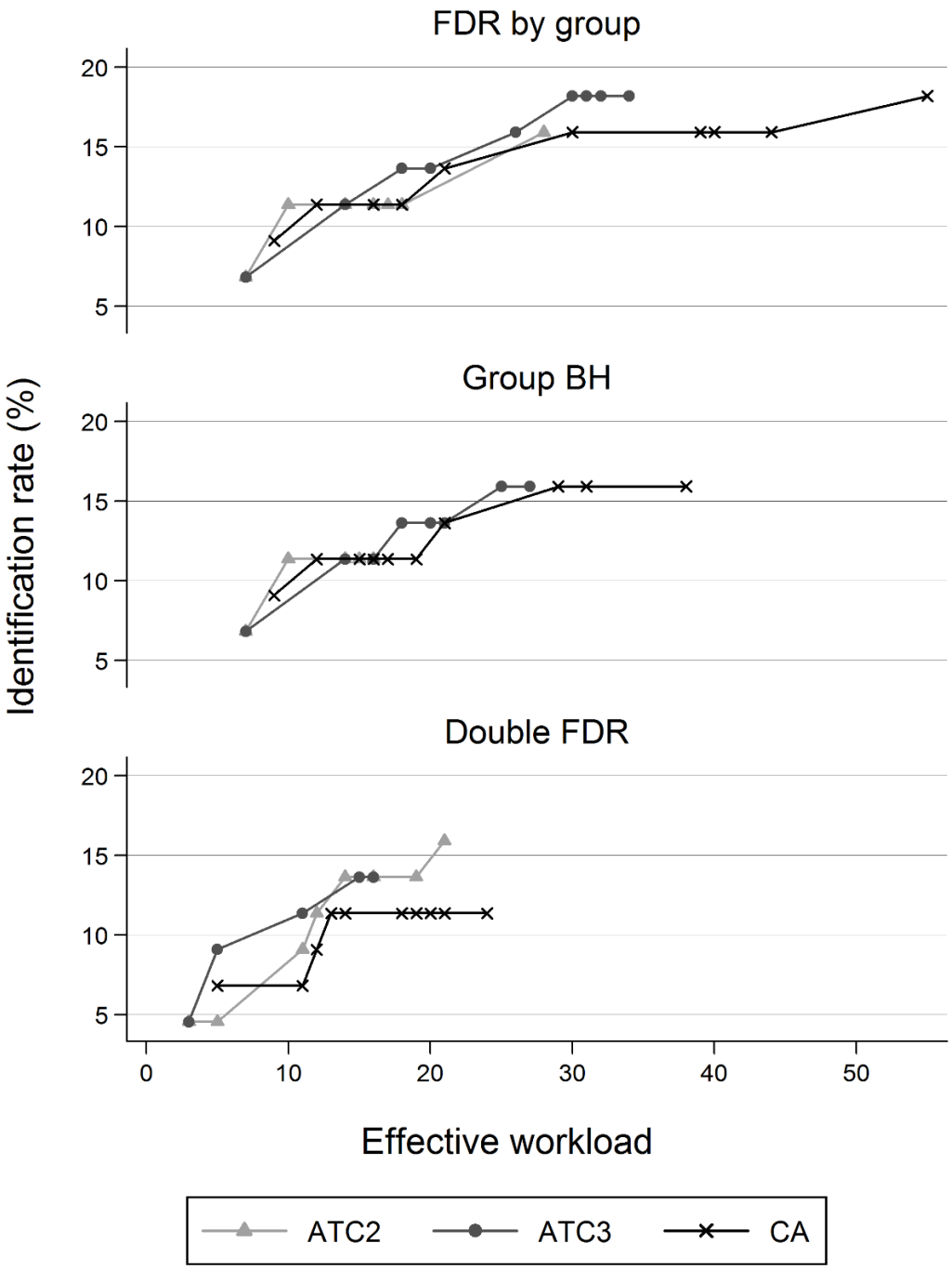


Figure 5.6. Identification rate (proportion of all "high risk" medications that are identified as signals) against effective workload (the total number of medication signals) for FDR by group, group BH and double FDR methods. Grouping is by ATC2, ATC3 codes and by CA, and each point corresponds to a different level of FDR for that type of grouping (in 5% increments from 5% to 50%).

5.3.3. Comparison of single and grouped FDR procedures using ATC3 codes to group medications

Figure 5.7 presents the number of signals identified in each risk category using ATC3 groupings for each of the four FDR methods across a range of FDR cut-offs from 5% to 50% (these numbers were also displayed in Table 5.1). Higher FDR cut-offs resulted in a greater total number and proportion of signals being identified for all methods. An FDR of 50% resulted in 4 to 5 times as many signals being identified as for an FDR of 5%. When considering the FDR cut-off of 50%, single FDR produced only 8 potential medication signals compared to 34 for the FDR by group, 27 for group BH and 16 for double FDR methods. Three (38%) potential signals identified by single FDR were in the "high risk" category compared with 8 (24%), 7 (26%) and 6 (38%) in the FDR by group, group BH and double FDR methods, respectively (Figure 5.7, Table 5.1). For single FDR 3 "high risk" signals were identified at an FDR cut-off of 15%; no further "high risk" medications were identified at higher choices of FDR cut-off for this method. Instead only one additional "low risk" medication was identified as a signal at an FDR cut-off of 35% and above. The number of "high risk" medications identified by grouped FDR methods was often close to the total number of signals identified by single FDR (Figure 5.7, Table 5.1). FDR by group and group BH methods generally identified an additional "high risk" medication signal compared to double FDR, but at the expense of a marked increase in the total effective workload.

Statistically significant associations that were not considered to be signals

Statistically significant associations with a medication-CA count of less than 3 or a PRR<1 (protective associations) were excluded from the list of potential signals. Table 5.2 shows the number of these associations that arose when using ATC3 groupings, for each of the four FDR methods. No combinations with less than 3 counts were identified as significant associations by any FDR method with a cut-off of 10% or less. With an FDR cut-off of 10% or higher, FDR by group provided the most associations based on small numbers, with 18 medication-CA combination associations having less than 3 counts for an FDR cut-off of 50%. Group BH identified around half as many associations with under 3 counts as FDR by group, and double FDR less again. Single FDR did not identify any combinations with less than 3 counts as being associations. Across all levels of FDR cut-off, single and double FDR each identified at most 2 combinations with a PRR<1. FDR by group and group BH identified slightly more protective associations, with a maximum of 8 and 5, respectively, when using a cut-off of 50%. The final columns in Table 5.2 show the total number of combinations and medications (i.e. effective workload; as displayed in Table 5.1) that were signals.

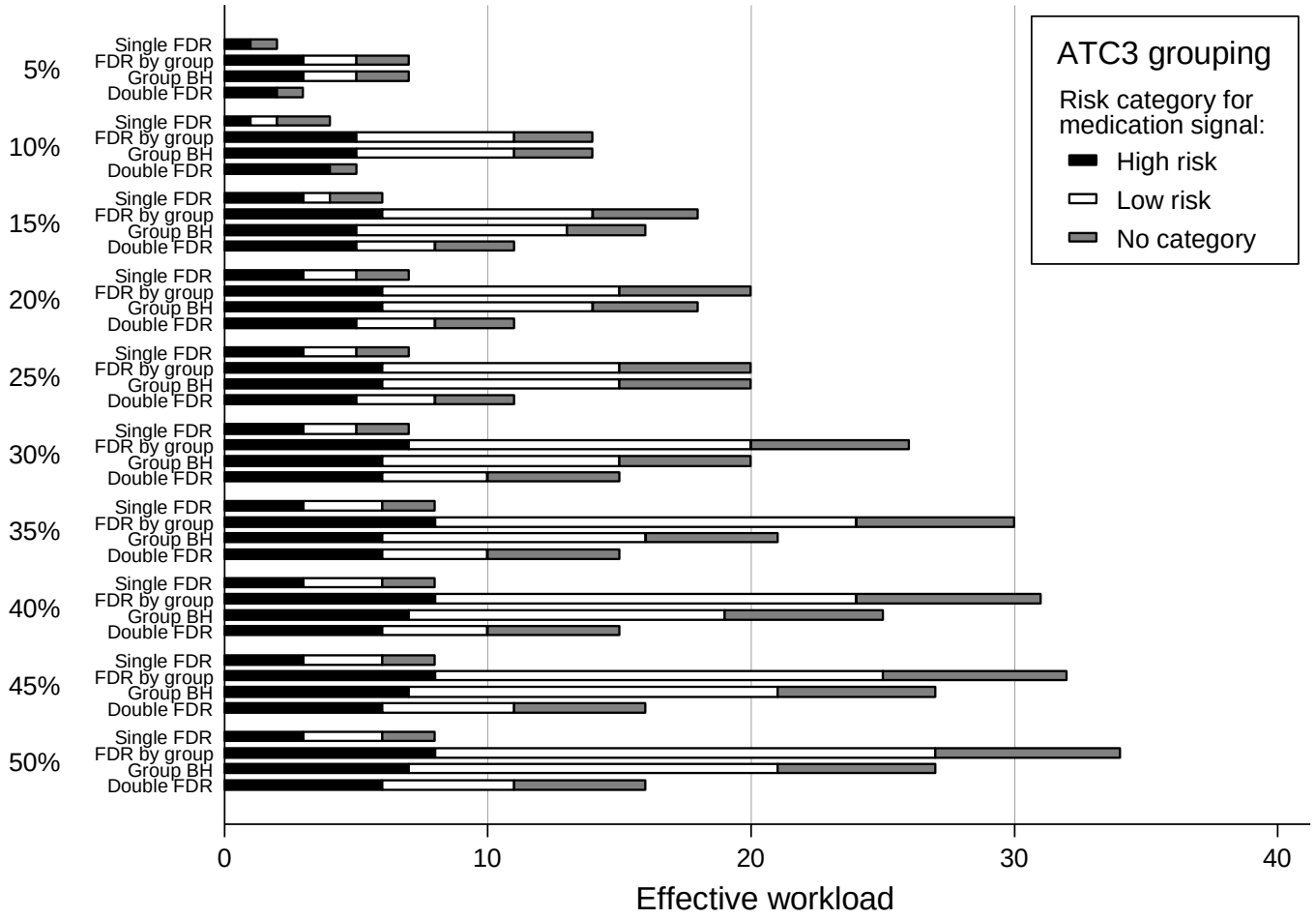


Figure 5.7. The number of signals detected in each risk category using ATC3 codes to group medication-CA combinations according to four FDR procedures, with FDR cut-offs ranging from 5% to 50%.

Table 5.2. Medication-CA combinations passing FDR adjustment but subsequently excluded from the set of potential signals due to low cell counts or protective associations, for all FDR methods and ATC3 groupings.

FDR cut-off level	FDR method	Combinations passing FDR adjustment	Exclusion due to:		Total signals after exclusions:	
			< 3 exposures	PRR <1	Medication-CA combinations	Medications [a]
5%	Single	2	-	-	2	2
	By group	12	-	2	10	7
	Group BH	12	-	2	10	7
	Double	5	-	-	5	3
10%	Single	4	-	-	4	3
	By group	19	-	2	17	14
	Group BH	19	-	2	17	14
	Double	7	-	-	7	5
20%	Single	9	-	1	8	7
	By group	32	4	2	26	20
	Group BH	28	4	2	22	18
	Double	13	-	-	13	11
30%	Single	10	-	1	9	7
	By group	50	7	4	39	26
	Group BH	33	4	2	27	20
	Double	21	1	1	19	15
40%	Single	12	-	2	10	8
	By group	67	12	6	49	31
	Group BH	46	7	3	36	25
	Double	21	1	1	19	15
50%	Single	12	-	2	10	8
	By group	80	18	8	54	34
	Group BH	53	8	5	40	27
	Double	29	2	2	25	16

[a] *This is the effective workload*

"High risk" proportion, identification rate and effective workload

Figure 5.8 shows the "high risk" proportion and identification rate for the four FDR methods using ATC3 groupings. The "high risk" proportion was lowest using the FDR by group and group BH methods, and decreased as the FDR cut-off level increased and more signals were identified overall. For all methods, the effective workload also increased as the identification rate increased. Double FDR gave the highest "high risk" proportion and identification rates per effective workload across most levels of FDR cut-off. The estimated identification rate was higher for group BH and FDR by group methods for levels of FDR cut-off that identified over 20 total medication signals.

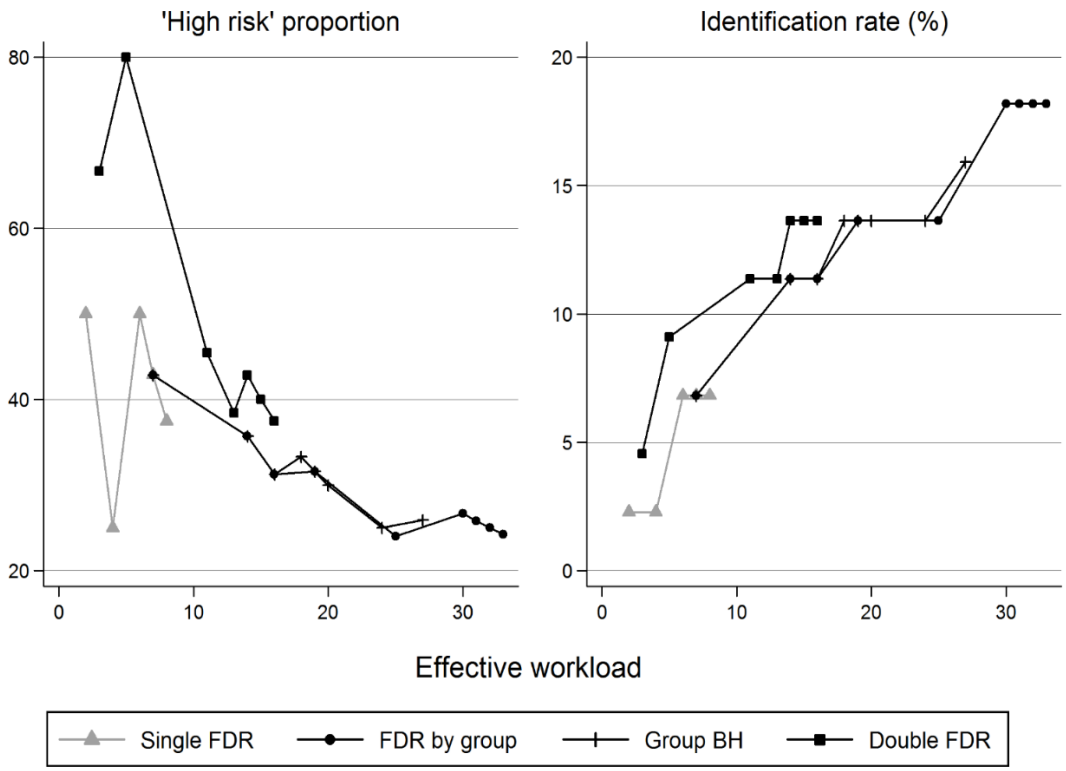


Figure 5.8. "High risk" proportion (percent of all medication signals that are in the "high risk" category; left panel) and identification rate (proportion of all "high risk" medications that are identified as signals; right panel) plotted against the effective workload (the total number of medication signals) for four FDR procedures using ATC3 groupings.

Another smileplot of P-values from all medication-CA combinations is displayed in Figure 5.9, now in which the black symbols highlight the 25 combinations (16 unique medications) identified as signals by double FDR with a cut-off of 50% and ATC3 groupings. Each data point in grey is a medication-CA combination that was not identified as a signal using double FDR with a cut-off of 50%, with different shaped markers used to show the assigned risk category of the medication in each combination. The data points above each dashed horizontal line correspond to medication-CA combinations that remain associations using that particular cut-off with the single FDR. There is only one combination in grey above the dashed line with a PRR>1 for the 50% cut-off. This combination was in the "low risk" category and is identified as a signal using single FDR with a cut-off of 50%, but was not a signal using the double FDR of 50%. Conversely, an additional 13 medications (16 combinations) shown in black below the FDR 50% dashed line were signals using double FDR 50% that were not signals using a single FDR of 50%. Of these, 4 belonged to the "high risk" category, 4 to the "low risk" category and 5 were not assigned a risk category.

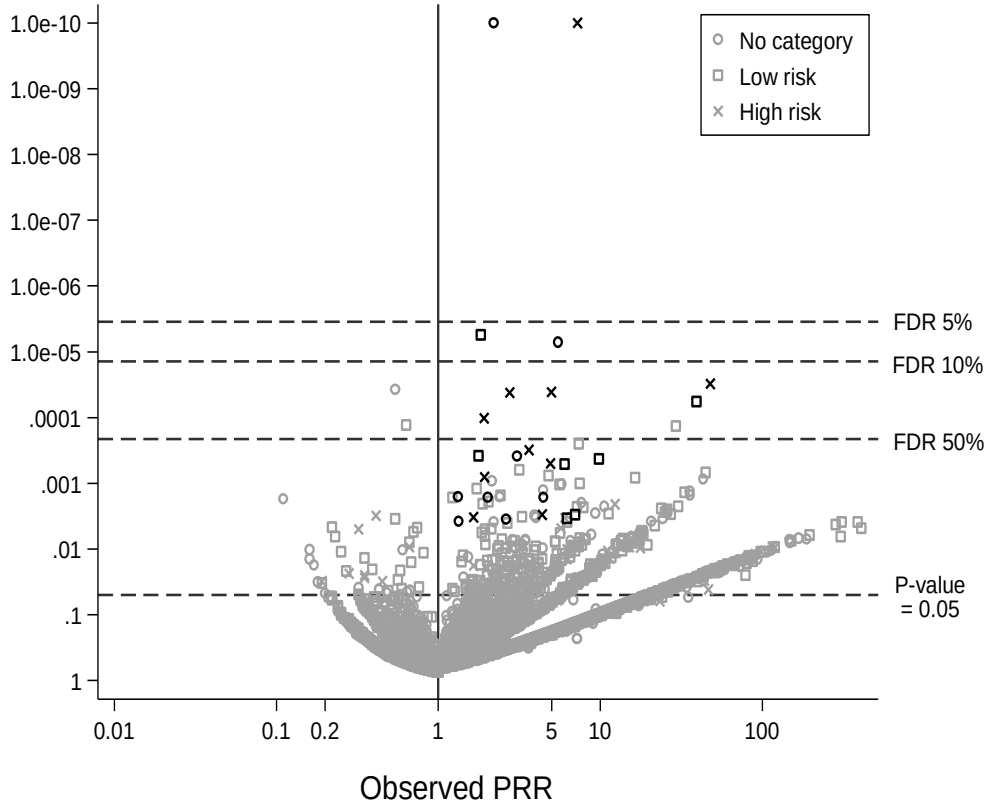


Figure 5.9. Smile plot of the observed PRR against the unadjusted P-value from Fisher's exact test for 28,396 medication-CA combinations, with different symbols according to Australian risk categories. Symbols in black indicate medication-CA combinations identified as potential signals using a double FDR with a cut-off of 50%. Single FDR cut-offs are indicated by dashed horizontal lines. Two P-values of 1.4e-17 and 2.0e-17 are shown at P=1.0e-10 for illustration purposes.

Table 5.3 shows the distribution of "high risk" combinations within each type of ATC grouping considered, including the additional grouping by ATC4 coding. The proportion of groups without any "high risk" medications was 9% higher in the ATC3 than ATC2 grouping, and 7% higher again when using ATC4 groupings. Note that when grouping medication-CA combinations by CA, the proportion of groups with no "high risk" medications was always non-zero and almost identical for every group, since each CA was tested once in combination with each medication. The only variation here comes from the fact that there was no data to perform Fisher's exact test for 369 combinations.

Table 5.3. Number of "high risk" medications per group according to different ATC groupings for medication-CA combinations.

Number of "high risk" medications in group	**ATC2** (3 digits)		**ATC3** (4 digits)		**ATC4** (5 digits)	
	N	%	N	%	N	%
0	44	72%	94	81%	215	88%
1	10	16%	13	11%	18	7%
2	2	3%	5	4%	7	3%
3	1	2%	3	3%	4	2%
4+	4	7%	1	1%	0	-
Total number of groups	61		116		244	

Table 5.4 provides details for the 16 medications (25 combinations) identified as signals by double FDR with a 50% cut-off and ATC3 groupings, as well as one additional medication signal identified by the single FDR 50%. Signals from double FDR included 5 antiepileptic medications (9 combinations), 4 insulin medications (7 combinations), 3 sex hormones (4 combinations), 2 antiasthmatic medications (3 combinations) one gynaecological medication (coded only to ATC4) and one medication for acid related disorders. Single FDR 50% identified only 7 of these medication signals over 10 combinations, including 2 of the antiepileptics (5 combinations), 2 antiasthmatics (2 combinations), one insulin-related medications and one sex hormone. Single FDR 50% also identified one further signal for the anxiolytic medication clobazam, which was not a signal with double FDR 50%.

Table 5.4. Summary of 26 medication-CA combinations identified as signals using single and double FDR procedures with a cut-off of 50%.

Medication Type	ATC3 group	ATC code	Risk cat.	Drug name	Congenital Anomaly	PRR (95% CI)	Unadj. P-value	N	Single/double FDR signal?
Acid related	A02B	A02BB01	X	Misoprostol	Anencephaly	47.8 (18.5, 123.6)	3.0E-05	3	Both
Insulin	A10A	A10AB01	-	Insulin (human)	Atrial septal defect	3.1 (1.9, 5.0)	3.9E-04	12	Double
		A10AB04	A	Insulin lispro	Patent ductus arteriosus	6.0 (2.8, 12.9)	5.1E-04	6	Double
		A10AB05	A	Insulin aspart	Lateral anomalies	7.0 (2.6, 18.7)	0.003	4	Double
					Ventricular septal defect	1.8 (1.3, 2.4)	3.8E-04	35	Double
		A10AC01	-	Insulin (human)	Atrial septal defect	2.0 (1.4, 3.0)	0.002	21	Double
					Unspecified CHD	2.6 (1.5, 4.6)	0.004	11	Double
					Patent ductus arteriosus	5.5 (3.1, 9.8)	7.1E-06	11	Both
Gynaecologic	G02C	G02CA	-	Sympathomimetic	Tricuspid atresia/stenosis	4.5 (1.9, 10.3)	0.002	8	Double
Sex Hormones	G03D	G03DA03	D	Hydroxyprogesterone	Atrial septal defect	1.9 (1.4, 2.7)	8.0E-04	27	Double
		G03DA04	-	Progesterone	Atrial septal defect	1.3 (1.1, 1.6)	0.002	127	Double
		G03DB01	-	Dydrogesterone	Atrial septal defect	2.2 (1.9, 2.6)	1.4E-17	150	Both
					Hypospadias	1.3 (1.1, 1.6)	0.004	94	Double
Antiepileptic	N03A	N03AA02	D	Phenobarbital	Cleft lip ± palate	4.9 (2.5, 9.7)	5.0E-04	7	Double
		N03AF01	D	Carbamazepine	Atrioventricular septal defect	4.4 (2.0, 9.8)	0.003	6	Double
					Spina Bifida	3.6 (2.0, 6.5)	3.1E-04	11	Double
		N03AG01	D	Valproic acid	Atrial septal defect	1.6 (1.2, 2.3)	0.003	34	Both
					Cleft palate	2.8 (1.8, 4.2)	4.2E-05	21	Both
					Hypospadias	1.9 (1.4, 2.6)	1.0E-04	38	Both
					Spina Bifida	7.2 (5.1, 10.2)	2.0E-17	32	Both
		N03AX09	D	Lamotrigine	Spina Bifida	5.0 (2.7, 9.0)	4.1E-05	10	Both
		N03AX14	B	Levetiracetam	Cleft palate	6.2 (2.6, 15.0)	0.003	4	Double
Anxiolytic	N05B	N05BA09	C	Clobazam	Atrioventricular septal defect	29.2 (11.0, 77.6)	1.4E-04	3	Single
Antiasthmatic	R03A	R03AC02	A	Salbutamol	Congenital hydronephrosis	1.8 (1.4, 2.3)	5.5E-06	66	Both
	R03C	R03CA02	A	Ephedrine	Congenital hydronephrosis	9.8 (4.8, 20.0)	4.2E-04	4	Double
					Multicystic renal dysplasia	39.4 (15.0, 103.3)	5.7E-05	3	Both

Table 5.5 displays the estimates, risk category and the FDR thresholds for single and double FDR for each of eight known medication-CA associations taken from the review by van Gelder et al. [2014], as used to validate signals in Luteijn et al. [2016] for the initial EUROmediCAT signal detection analysis. There were no exposures for combination of cleft lip ± palate with oxprenolol, hence this was not included in the analysis. The combinations of naproxen and phenytoin with cleft lip ± palate had only zero and one exposures, respectively, hence these PRRs were very imprecisely estimated. The combination of progesterone with hypospadias was present in the data and had a PRR>1, but was not a statistically significant association. Valproic acid in combination with both atrial septal defect and cleft palate were picked up by double FDR at lower cut-offs than for single FDR (for atrial septal defect this was not a signal for any single FDR cut-offs lower than 50%). Single and double FDR both picked up the signals for valproic acid with spina bifida at the same thresholds, however neither method retained the combination of valproic acid and craniosyntosis as a signal after FDR adjustments at a cut-off of 50% or less.

Sensitivity analysis: medications not assigned a risk category

Of the 523 medications in the EUROmediCAT data for these analyses, 35% were not present in the Australian classification system database and so these were not assigned a risk category (see section 4.4.3 and Table 4.9). Sensitivity analyses were therefore performed to check the effect on FDR results according to a number of hypothetical situations. Firstly, medications without a risk category were assumed to either be all “high risk” or all “low risk”, representing the two most extreme situations. The effect of assuming that a medication without a known risk category was “high risk” if there was at least one other “high risk” medication in that group, and “low risk” if all medications in the same group were either “low risk” or did not have a risk category assigned was also investigated. For these three situations, comparisons between the different FDR methods were the same as described previously, and all conclusions in this chapter remained unchanged (data not shown).

Table 5.5. Evaluation of methodology as done for previous EUROmediCAT signal detection, using selected known medication–CA associations identified by van Gelder et al [2014].

ATC	Medication	Congenital anomaly	Exposed foetuses	Exposed foetuses with specified CA	PRR (95% CI)	P-value	Single FDR	Double FDR [a]	Risk Category
G02CC02	Naproxen	Cleft lip ± palate	3	0	0 (0 – 25.13)	0.86	-	-	C
C07AA02	Oxprenolol	Cleft lip ± palate	0	0	NA	NA	-	-	C
N03AB02	Phenytoin	Cleft lip ± palate	17	1	1.15 (0.03 - 7.45)	0.59	-	-	D
G03DA04	Progesterone	Hypospadias	1108	106	1.14 (0.92 - 1.41)	0.12	-	-	NA
N03AG01	Valproic acid	Atrial septal defect	235	34	1.75 (1.17 - 2.54)	0.003	>50%	**<45%**	D
N03AG01	Valproic acid	Cleft palate	235	21	2.92 (1.75 - 4.63)	0.00004	<20%	**<5%**	D
N03AG01	Valproic acid	Craniosynostosis	235	6	3.06 (1.09 - 6.94)	0.017	>50%	>50%	D
N03AG01	Valproic acid	Spina bifida	235	32	8.19 (5.35 - 12.18)	2.0E-17	<1%	<1%	D

[a] *Using ATC3 groupings*

5.4. Discussion

In this chapter, methodology from the recent EUROmediCAT signal detection analyses was compared to methods of multiple testing adjustment that incorporated prior information about the grouping of medications or CAs into the analyses when determining the statistical significance of each medication-CA combination.

5.4.1. Comparing different groupings of medication-CA combinations

Three different types of grouping for the medication-CA combinations were first assessed, and grouping by ATC3 codes was found to have slightly better performance in terms of proportion of "high risk" signals being identified as well as the balance between identification rate, "high risk" proportion and effective workload. The proportion of groups containing no "high risk" medications was highest when using ATC4 groupings and lowest using ATC2 groupings (Table 5.3).

Groups with a greater number of "high risk" medications should be expected to contain more signals; in the ATC3 and ATC4 groupings it was therefore expected that more groups would be "thrown out" at the first stage of the grouped FDR methods compared to use of ATC2 groups, since there were less groups including a "high risk" signal. If a greater number of groups are excluded after this first stage, then the second stages of the grouped FDR procedures are done across smaller numbers of medication-CA combinations, leading to less strict adjustment to individual P-values (as a smaller number of tests requires less stringent adjustment to achieve the required level of FDR control). This may result in more signals being picked up for groups where there is at least one signal, although this does depend on which particular groups are excluded because each group does not have the same number of medications. For example, there was an average 242 combinations per ATC3 group with a minimum of 55 (i.e. one medication in combination with 55 CAs) and a maximum of 1,071 (i.e. 20 medications in combination with 55 CAs, excluding some combinations for which there was no data to perform Fisher's exact test - see section 5.3.2). On the other hand, a greater number of groups also means a stricter threshold in practice for the first stage of the grouped FDR procedures. This is because the adjustments in the first step are done across the set of P-values made up of one representative P-value from each group, and a greater number of tests requires more stringent control to achieve the same level of FDR. However, the double FDR (for example) calculates the representative P-values using a single FDR adjustment within each group to obtain a minimum FDR adjusted P-value, which will be less strict for each individual group if the

overall number of groups is larger (since each group size will be smaller). These two factors may have balanced the effect of varying the number of groups on how stringent the FDR adjustment to the P-values is in the first stage of the grouped FDR procedures. This means, for example, that a group which has one signal that is statistically significant at a level of adjustment using the 116 ATC3 groups may have been excluded in the first stage of double FDR when using the 244 ATC4 groups since there will be more than twice as many P-values in this multiple correcting adjustment. As a sensitivity analysis, therefore, FDR procedures were repeated using the ATC4 codes as the grouping; however, this did not offer improved performance for any of the FDR methods considered.

Grouping by CA resulted in 55 distinct groups with an average of 518 (range 325-523) unique CA-ATC5 medication combinations per group previously (see section 5.3.2). In the previous chapter, Table 4.9 also showed that 44 of the 523 unique medications in these analyses were in the "high risk" group category. When grouping combinations by CA, therefore, the maximum possible number of "high risk" medications per group was 44, and this was the case for 32 (58%) of the groups in the analysis. A further 20 (36%) CAs contained 43 of the "high risk" medications and only three subgroups contained fewer: 30/325 combinations for the conjoined twins subgroup, 41/501 for lateral anomalies and 42/504 for congenital glaucoma. Grouping using the CAs for FDR adjustment did not perform better than the groupings by ATC codes, which might be influenced by the use of the Australian classifications to judge the methods, since these are not specific to the type of CA.

To summarise, the type of groupings used in this chapter did not give markedly different results and so it is unlikely that further types of grouping based on available ATC or EUROCAT coding structures would provide a more useful comparison of the methods assessed. For double FDR, however, ATC3 grouping provided the highest proportion of signals being in the "high risk" category, with a similar effective workload as the other types of grouping (see Figure 5.4, Figure 5.5 and Figure 5.6). The ATC3 groupings were therefore used to directly compare the different FDR methods.

5.4.2. Comparison of FDR procedures grouping combinations using ATC3 level codes

When using ATC3 groupings the double FDR performed better than the other two methods that grouped medication-CA combinations, and use of this method is therefore recommended in practice when considering groups in signal detection analyses for CA data. Group BH and FDR by group both identified one additional "high risk" medication signal

compared to double FDR for most levels of FDR cut-off; however, this was at the expense of a disproportionate increase in the total effective workload. Double FDR showed the best "high risk" proportion and identification rate across all levels of FDR cut-off. FDR by group and group BH only performed better than double FDR in terms of identification rate for larger numbers of medications (i.e. at least 20) being identified as signals. A considerably larger number of total medication signals for follow up were required to achieve these higher identification rates, representing a potentially unrealistic workload, of which less than 30% may be expected to be "high risk" signals (Figure 5.8).

Choice of FDR cut-off threshold

Another variable to consider in FDR methods is the choice of FDR cut-off level α, which here was allowed to range in 5% increments from 5-50%. The choice of FDR cut-of aims to balance the proportion of false negative and false positive associations. If this value is too high, the resulting workload for follow up investigation of associations may be larger than is acceptable, depending on available resources. There may also be potential for unwarranted anxiety to be caused for pregnant women if false positive associations are reported for medications that they may have to take, even if the resulting investigations are inconclusive. A very low FDR cut-off, conversely, may miss important signals and result in a delay of detecting a teratogen until more data (i.e. more women being exposed to a potentially harmful medication) are available. The cut-off value for the FDR should therefore be re-evaluated frequently, and in practice this will depend on the resources available to investigators for follow-up of signals. It is important also to remember that signal detection is only the first step in the process of safety surveillance for medication use in pregnancy. Results from signal detection are subject to detailed evaluation to determine which medications require independent confirmation or further investigation [Given et al., 2016]. EUROmediCAT also separately considers four medication classes (new antiepileptics, insulin analogues, antiasthmatics, and antidepressants/SSRIs) in more detailed analyses, with an emphasis on hypothesis testing rather than hypothesis generating [de Jong-van den Berg et al., 2011]. When combined, these different approaches to signal detection limit should the consequences of setting an FDR cut-off that is too strict, because many resulting signals are likely to already been investigated or be under investigation elsewhere. In these more detailed later stages of the signal detection process further factors are also considered, including data quality, registry-specific queries, consistency with other literature, co-occurring medication exposures and biological plausibility.

The actual value of the cut-off chosen for the FDR represents the estimated maximum proportion of observed associations that are likely to have arisen due to chance; an FDR cut-off of 50% means that up to 50% of the signals identified are expected to be false positive associations, i.e. as many as half of the observed associations are likely to have arisen by chance. Note that this is a percentage of the associations identified by the method and does not take into consideration those associations that were not included in the resulting set of signals after exclusions, e.g. 30 total combinations were identified as associations after double FDR 50% with ATC3 groupings, of which only 25 were identified as signals after exclusions based on low frequency and PRR<1 (Table 5.2). Associations with low cell counts (<3) are excluded specifically because evidence from such a small number of cases is judged to be insufficient, and these associations may therefore be more likely to have arisen due to chance i.e. to be the expected false positives. In general, then, when using FDR methods for signal detection a smaller proportion than $\alpha\%$ of the resulting signals may therefore arise due to chance alone. As such, the cut-off choice of α for the FDR methods is not a clear indication of how many of the signals are likely to be true associations. In these analyses, the "high risk" proportion was also evaluated, which aimed to estimate the proportion of signals that are likely to be "true associations" (i.e. because these are already known to be "high risk" medications for CAs). Figure 5.8 showed that the proportion of medication signals known be "high risk" decreased as the FDR cut-off α increased; with a higher α, a greater proportion of signals are expected to be false positives and a lower proportion of "high risk" medications is therefore observed.

The group BH method, as described by Hu et al. [2010], was of interest for these analyses because it is aimed specifically for situations where there are thousands of P-values, as is the case in analysis of EUROmediCAT data. However, the FDR control for this method at the target level of α is maintained only in an asymptotic sense i.e. for very large numbers of tests. All groupings considered here provided very high estimates $\hat{\pi}_0$ of the average proportion of null hypothesis across the groups, reflecting the fact that the majority of groups do not contain any signals (i.e. the majority of groups have only true null hypotheses), such that most types of medication taken are not teratogenic. For all medication-CA combinations in this data, $\hat{\pi}_0$ was estimated to be 0.9992, 0.9989 and 0.9992 for ATC2, ATC3 and CA groupings, respectively. Hu et al. [2010] noted that the group BH method performs best in situations where the estimate $\hat{\pi}_{0i}$ of the proportion of null hypotheses in each group significantly differs across the groups. This was not the case for the EUROmediCAT data, since over 70% of groups had an estimated zero proportion of

rejected null hypotheses for all types of grouping and level of FDR cut-off. For those groups where $\hat{\pi}_{0i} < 1$, the remaining estimates of $\hat{\pi}_{0i}$ were around 0.98 on average, ranging from 0.84 to 0.99 across all groupings and FDR cut-off levels. This may explain why the group BH method did not perform particularly well for these data.

For the FDR by group method, the choice of cut-off level $\alpha_1$for the first stage of the FDR process can fail to control the overall FDR rate at $\alpha\ (= 2\alpha_1)$. A simulation study by Mehrotra and Adewale [2012] indicated that the actual FDR associated with this approach can in fact be up to 2-3 times larger than the target overall FDR level in some situations. In practice, the groups that dropped out after the first stage of this process were unlikely to include any potential signals after FDR adjustment in the second stage, even at the increased cut-off level of $\alpha_2 = 2\alpha_1$. This procedure therefore provided very similar results as would be obtained from a separate FDR procedure for each group, i.e. not excluding any groups first, as is done here. Note that FDR adjustments at a level α applied separately to a number of groups does not necessarily imply FDR control at level α overall (i.e. across the whole study) [Benjamini and Yekutieli, 2005].

Despite these considerations surrounding the choice of FDR cut-off, the focus for these analyses was to compare workload and identification rate for the methods explored, and hence the actual level of FDR control was not a key concern. More important is to achieve the highest possible detection rate (here estimated using the identification rate) for a workload that would be deemed acceptable for follow up of signals in practice. The choice of FDR cut-off for any FDR method should therefore be considered flexible and be adjusted accordingly to reflect the available resources for follow up of signals of any particular project.

5.4.3. Signals identified by single and double FDR 50% with ATC3 groupings

Signals identified by FDR methods in the context of evidence from other studies

Further evaluation of associations identified in a signal detection process was not an aim of this thesis; however, this section will briefly refer to the existing evidence and knowledge regarding the 17 medications in 26 combinations identified as signals by the single and double FDR methods with a cut-off of 50% (as displayed in Table 5.4) in order to provide some context for these associations. The most common type of signals were in the group of antiepileptics; these are well-established as being a teratogenic group of medications [Bruni and Willmore, 1979, Petersen et al., 2017, Veroniki et al., 2017]. The next most common type was insulin, the use of which during pregnancy is a marker of maternal diabetes. It is

also well-recognised that it is poor control of maternal hyperglycaemia that causes an increased risk of malformations, rather than the insulin itself [Allen et al., 2007, Zabihi and Loeken, 2010, Charlton R. A. et al., 2016, de Jong et al., 2016b]. An increased risk of specific CAs has also been demonstrated following first trimester exposure to some antiasthmatics [Blais et al., 2010, Lim et al., 2011]. Progesterone and its derivatives have been linked to an increased risk of hypospadias [Carmichael et al., 2005] and CHDs [Zaqout et al., 2015]. Misoprostol is the only medication in Table 5.4 that is in the highest risk category X, indicating there is a high risk that it causes permanent damage to the foetus. This medication has various uses including ulcer prevention, labour or abortion induction and the treatment of postpartum bleeding, and has been shown to cause birth defects such as brainstem injuries [Vauzelle et al., 2013]. The combination of the ATC4 coded sympathomimetic medication G02CA with tricuspid atresia and stenosis was unexpected in that a labour suppressing medication is not usually taken as early in pregnancy as the first trimester. Exposures to this particular combination were present only in the two Italian registries, Emilia-Romagna (n=2) and Tuscany (n=6). Across the whole dataset, however, there were 583 exposures with the ATC4 level code G02CA, and 97% of these were in Emilia-Romagna (n=210) or Tuscany (n=357). Further exposures to more specific ATC5 codes G02CA01 (n=232) and G02CA03 (n=68) were reported, meaning exposures to a G02CA medication were present in seven registries: Emilia-Romagna, Tuscany, Zagreb, Poland excluding Wielkopolska, Poland Wielkopolska, Antwerp and Vaud. Exposures to G02CA medications in the Polish registries and Vaud were specified as labour suppressants taken in the first trimester, according to additional notes fields. For cases in the Tuscany registry this medication was generally named as Vasosuprina; no specific medication name was specified for Emilia-Romagna. The efficacy of Vasosuprina's active ingredient (isoxsuprine hydrochloride) as a treatment for women at risk of abortion or preterm labour has been evaluated in a number of (mostly Italian) studies [reviewed by Giorgino and Egan, 2010]. This medication was taken at all gestational ages, including the first trimester in almost half the studies assessed in this review, and it therefore seems reasonable to assume that it was indeed taken in the first trimester in practice here, particularly given that there are a considerable total number of exposures to this medication in the data. However, although first trimester medication use is a pre-requisite for inclusion of a case in this data, records often lack detailed information regarding the specific timing of medication use. It is therefore also possible that these medications were taken at a later stage of pregnancy. Queries such as this regarding cases included in the analysis of signals

would be followed up with the specific registries as part of the next stage of the signal management process, which can help determine whether there is a possibility of misreporting or data errors. Another thing to note about this combination is that the reason it was a signal after double FDR only (and not after a single FDR adjustment) is because G02CA belongs to the ATC3 group G02C "other gynaecological" group of medications, in the group of genitourinary system and sex hormone medications. This ATC3 group passed the first stage of the double FDR adjustment due to the small P-value (P=0.00004) of the medication G02CA in combination with atrial septal defect, another CHD. The combination of G02CA with atrial septal defect is in fact a protective association (see page 170 for further discussion of protective associations), but it is of note that this medication is the only association in Table 5.4 for which there is not already a combination in the same ATC3 group that was also a signal when using a single FDR. This highlights how P-values of combinations in groups including at least one signal (i.e. in groups that pass the first stage of the double FDR) are penalised less strictly by double FDR than single FDR, because the adjustment is done on a smaller set of combinations overall after discarding information from all groups not passing the first stage of the double FDR.

There was only one combination that was a signal using single FDR but not double FDR, for the CHD atrioventricular septal defect in combination with the anxiolytic medication clobazam (N05BA09). This medication was also a signal in the EUROmediCAT signal detection analysis of Luteijn et al. [2016], but was not included for follow-up in Given et al. [2016] because the anxiolytics are a therapeutic subgroup of the psycholeptic medications and these were included in the separate EUROmediCAT work package for analysis of selective serotonin reuptake inhibitors, since these two classes of medications are frequently co-prescribed [EUROmediCAT, 2011]. An excess risk of the CHD Ebstein's anomaly was reported for the group of psycholeptic medications in this study, but it was noted that there was limited staticstical power to assess these medications; the reported association was based on small numbers and hence requires further follow-up [EUROmediCAT, 2011].

Comparison of signals from single and double FDR in this chapter with results from previous EUROmediCAT signal detection analyses

The previous EUROmediCAT analysis identified 39 signals for future follow-up [Luteijn et al., 2016]. As part of the EUROmediCAT project, any signals belonging to four groups of medications were already being separately investigated: insulin/insulin analogues [de Jong et al., 2016b], antiasthmatic medications [Garne et al., 2015], antiepileptic medications

[Charlton R. et al., 2015, de Jong et al., 2016a] and psycholeptic medications / selective serotonin reuptake inhibitors [Wemakor et al., 2015]. Of the 39 signals identified by Luteijn et al, 14 were signals for insulin/insulin analogues, 4 for anti-asthmatic medications, 8 for antiepileptic medications, 2 for selective serotonin reuptake inhibitors / psycholeptic medications. There were 11 signals of other types of medications, which were then followed up in more detail in a separate study as the next stage of the signal management process [Given et al., 2016]. When considering these 39 signal combinations in the analyses done for this thesis:

- 21 of these combinations were signals, of which
  - 10 combinations were also signals in these analyses: 3 after double FDR only, 1 after single FDR only, and 6 after both a single and a double FDR procedure
  - 8 combinations were with ATC4 coded medications, but for which there was a signal of a more specific ATC5 medication with the same CA in the current analyses
  - 3 combinations were with the aggregate subgroup of CHDs, but for which there was a signal in the analysis of a more specific CHD subgroup with the same ATC code in the current analyses
- 13 of these combinations were not present in the current analysis dataset, of which
  - 5 were medications in combination with an aggregate subgroup of CHD or severe CHD, which were not included separately in these analyses
  - 8 were ATC4 coded medications that were not included separately in these analyses (since a more detailed ATC5 code was available)
- 5 of these combinations did not reach statistical significance to be judged signals in these analyses

For the latter point, some of these are likely to be due to the smaller sample, as there were smaller numbers for many of the same medication-CA combinations in the dataset for this thesis. On the other hand, some combinations may have dropped out of the set of signals because the level of P-value cut-off differs by ATC3 group when using double FDR (and this will be lower than the overall P-value cut-off of the single FDR for some groups).

Of the 26 medication-CA combinations displayed in Table 5.4, 6 were not signals in Luteijn et al, where their P-values were slightly above the FDR-adjusted cut-off for statistical significance. Ten of the remaining combinations in Table 5.4 were signals in Luteijn et al, of which one combination (atrioventricular septal defect with clobazam) was a signal using

single FDR (in both Luteijn et al. and in this thesis), but not when using double FDR. For ATC3 groupings, the P-value for this combination was the minimum representative P-value in its group N05B, which included 9 different medications in this data. However, the N05B group was not included in the second stage of the double FDR procedure as it was dropped in the first stage (after FDR adjustment across the representative P-values from each of the 116 ATC3 groups). This highlights the way in which the double FDR procedure can result in a stricter adjusted P-value threshold for certain groups compared to the single FDR procedure. A further 6 combinations in Table 5.4 included the same medication but were only signals in Luteijn et al. only in combination with the aggregate subgroup of CHDs (i.e. being a signal with a more specific CHD in this chapter). Three combinations were a signal for the ATC4 code with the same CA (i.e. being a signal with a more specific ATC5 code in this chapter).

Comparison with other existing studies of teratogens

Table 5.5 highlighted the results from single and double FDR for eight known medication-CA associations taken from a review by van Gelder et al. [2014]. This review was a comprehensive synthesis of evidence regarding teratogens in scientific literature up to 2013, addressing the lack of evidence regarding teratogenicity of medications. The eight signals presented in Table 5.5 are reported associations between an ATC4 or an ATC5 coded medication and a specific CA, including those observed in case-control studies that were confirmed in at least two studies (and not refuted in any other studies). Whilst this list of eight medications is not comprehensive and does not consider associations of any newer medications that have appeared on the market since 2013, it is an objective set of known associations and was therefore used to evaluate the EUROmediCAT signal detection method [Luteijn et al., 2016]. Note that Table 5.5 shows similar estimates to those in Table 4 of Luteijn et al. Results for this small set of combinations indicate that double FDR performed slightly better than single FDR, in that two of the combinations were identified as signals by lower FDR cut-off levels. However, even in this small selection of combinations for which there was considered to be good evidence of an association, there were insufficient cases in the data for 3/8 of the combinations to even potentially be included in the resulting set of signals (less than 3 exposures to the specific medication-CA). This raises the point that known teratogenic medications may be underrepresented in the data, since they are rarely or no longer used in practice. Published studies documenting the use of oxprenolol, for example, have all been based on data prior to 1997 [Yakoob et al., 2013]. Furthermore, half of the 8 associations in this set relate to only one medication. A small set

of known associations such as this is therefore limited in its ability to judge a signal detection method. An approach that evaluates combinations across the whole dataset should be less prone to this issue, and this is why the Australian categorisation system was used here to try and identify "high risk" medications across the whole dataset.

Protective associations identified by single and double FDR procedures

A number of combinations with PRR<1 remained associations after the FDR procedures, indicating a "protective" association of a medication for a particular CA. Such associations are likely to be due to chance (i.e. up to 50% of observed associations are expected to be due to chance with an FDR cut-off of 50%) or due to biases as a result of the case-malformed control study design, where the controls all have at least one CA and one medication exposure. The antiepileptic medication lamotrigine (N03AX09), for example, was a signal for spina bfida in this dataset (Table 5.4); when the association between lamotrigine and different CA is examined, a large proportion of the controls are likely to be associated with lamotrigine via their association with spina bifida. A protective effect may therefore be observed for other CAs due to the large numbers of controls that are associated with lamotrigine. There was a significant protective association for the CHD ventricular septal defect in combination with lamotrigine (n=7) when using double FDR, but this association was not significant in analyses using single FDR. There were also protective associations for two medications G02CA (labour repressants, n=34) and N02BE01 (paracetamol, n=55) with the CHD atrial septal defect. It should again be noted that protective here is in comparison to other CAs and medications in the database, and does not imply that the overall risk of a CA is lowered by such a medication.

5.4.4. Use of risk categories to compare signal detection methods

Various drawbacks regarding the use of the risk categorisation system have been discussed previously (see chapter 4). The teratogenic risk of a medication is often specific to certain CAs, rather than there being an increased risk of malformations in general [Mitchell, 2016]. A key issue for these analyses is that there is no such differentiation between the categorised risks of each medication in terms of different CAs, since specific CAs are not taken into account in the Australian risk categorisation system. This may have an effect on the "high risk" proportion and identification rates assessed in this chapter, since these measures therefore do not differentiate between a medication associated with only one CA and one associated with a number of different CAs. For example, the medication valproic acid is associated with four CAs whilst the medication Salbutamol is associated with only

congenital hydronephrosis (Table 5.4); however both medications contribute only once each to the overall “high risk” proportion and to the identification rate.

Of the 16 medication signals identified by double FDR 50% 6 were “high risk”, 5 “low risk” and 5 were not assigned a risk category (Table 5.4). Due to 5 medications here having unknown risk categorisation, the proportion of medication signals that are likely to be “high risk” may be either under- or overestimated. On one hand, it may be assumed that if a medication is not given a risk category in the Australia categorisation system it is more likely to be of low risk, assuming that the harmful medications have already been identified. However, this would not be the case for newer and/or rarer medications, which are primarily what signal detection methods aim to identify. In fact, 35% of all medications included in the analysis were not assigned a risk category, and it is possible that the “high risk” proportion and identification rate have been underestimated. Two of the medications in Table 5.4 listed as “low risk” category A medications were insulin medications. It is not surprising to find the insulin medications in the low risk category according to the Australian classification system because, although the association between insulin medications and increased risk of CHDs is well known, this is often a result of the fact that the mother has diabetes, rather than the insulin medications being a risk factor in themselves. So whilst it is recognised that the insulin medications are often unlikely to be teratogenic themselves, it is important that they are appearing as signals in any signal detection method, and the risk categorisation system will not reflect this. In Table 5.5, two of the medications that were deemed to show evidence of being teratogenic in van Gelder et al. (naproxen and oxprenolol) are assigned as “low risk” category C medications according to the Australian risk categorisation system. This highlights the fact that the assigning of medications to the categories in these types of systems can be based on differing expert opinions and assessments of the available evidence regarding the potential risks of each medication for use in pregnancy. A third medication in Table 5.5, Progesterone, was not assigned a risk category at all, although this medication can be found in the same group of codes as G03DA02 (medroxyprogesterone) and G03DA03 (hydroxyprogesterone), which are both “high risk” category D medications.

5.4.5. Summary and conclusions

In summary, this chapter showed that the double FDR using ATC3 groupings performed better than other methods considering grouping, including the currently used single FDR procedure. Difficulties in the comparison of these methods was discussed, specifically the ability to judge the absolute strengths of any signal detection method due to issues with their validation. However, the double FDR was judged to be superior based on the findings and metrics used in these analyses, and this method can also be easily implemented in practice (the Stata code used to run a double FDR procedure with ATC3 groupings is presented in Appendix B2). However, FDR procedures presented in this chapter considered grouping of medications and/or CAs at the point of determining the statistical significance following a separate statistical test of each medication-CA combination. The next chapter considers the use of BHMs to directly model potential group effects in EUROmediCAT data, in order to determine whether such models provide better results than FDR procedures for signal detection in CA data.

Chapter 6: Analysis of EUROmediCAT safety of medication use during pregnancy II: Bayesian hierarchical models

6.1. Introduction

In this chapter, methodology from DuMouchel [1999], Berry and Berry [2004] and Brook [2011] is applied to EUROmediCAT data to investigate whether use of BHMs that group together medications and/or CAs can improve signal detection methods for CA data. These models were developed in the context of signal detection in SR databases or for clinical trials of experimental pharmaceutical medications, but have not previously been applied to CA data. Results are compared to those obtained using a single FDR, as this is the method currently used by EUROmediCAT for signal detection purposes, as well as the double FDR method with ATC3 medication groupings, which showed improvement over the single FDR in Chapter 5.

6.2. Methods

Signal detection analyses in this chapter were investigated using BHMs that take groupings of medications (using ATC3 codes) and/or CAs (using type of CA grouped by their organ system classes) into account. Individual Bayesian models not incorporating any groupings were also analysed for comparison purposes. The methodology in this chapter combines the Gamma Poisson Shrinker and a BHM (see section 4.2.2 for further details) to apply to a large database containing many cells determined by combinations of a specific medication and a specific CA.

Calculation of expected values

As described in Table 4.1, the PRR was defined as

$$PRR_{ij} = \frac{c_{ij} \,/\, c_{i.}}{c_{i'j} \,/\, c_{i'.}}$$

Assuming no association between i and j, the expected count for c_{ij} was calculated using the marginal totals for medication i and CA j

$$E_{ij} = \frac{c_{i.}\, c_{.j}}{N}$$

Under assumption of independence between the medications and the CAs, the PRR was expressed as the ratio of the observed to expected counts

$$PRR_{ij} = \frac{c_{ij}}{E_{ij}}$$

To illustrate this, a hypothetical 2-by-2 table of counts for the combination of a medication i with a CA j is displayed in Table 6.1. In this example, there are a total of 1,250 exposures (N) in which the observed marginal count for medication i is 140 ($c_{i.}$) and the observed marginal count for CA j is 150 ($c_{.j}$). Under the assumption that medication i and CA j are independent, 12% (=150/1250) of exposures to medication i would be expected to also have CA j, and the expected count for this particular combination would then be 17 exposures, i.e. 12% of 140 or $\frac{140x150}{1{,}250}$. When comparing the observed and the expected count in this example, 30 is therefore being compared to 17 such that 30/17 = 1.76 implies that the observed combination was 76% more likely than expected if medication i and CA j were independent.

Table 6.1. Example of a hypothetical 2x2 table for analysis of the relationship between medication i and congenital anomaly j.

	Cases: Foetuses with CA j	**Malformed controls:** Foetuses with CAs other than j	Total
Exposed to medication i	$c_{ij} = 30$	$c_{ij'} = 110$	$c_{i.} = 140$
Unexposed to i, but exposed to at least one other medication in the data	$c_{i'j} = 120$	$c_{i'j'} = 990$	$c_{i'.} = 1{,}110$
Total	$c_{.j} = 150$	$c_{.j'} = 1{,}100$	$N = 1{,}250$

Note that the above definition for E_{ij} means that the count c_{ij} for each individual combination is essentially contributing twice to the model, since it is counted in the marginal totals for both the medication and the CA. As a sensitivity analysis, therefore, the expected counts were also calculated using an alternative definition to assess whether the expected number of exposures for each medication-CA combination under the assumption of independence between medication i and CA j may be better estimated. In this alternative definition, the total number of exposures, the probability of exposure to medication i independent of CA j, and the probability of CA j independent of medication i were multiplied together as follows

$$Alternative\ E_{ij} = N \times \frac{c_{ij'}}{c_{.j'}} \times \frac{c_{i'j}}{c_{i'.}}$$

6.2.1. Data structure for Bayesian models according to different types of grouping for medications and congenital anomalies

Separate Bayesian models for each medication-congenital anomaly combination

Firstly, a separate Bayesian model was applied to each medication-CA combination, without considering any groupings of medications or CAs. This data structure is displayed in Table 6.2; in this setting, each count c_{ij} of a particular medication i in combination with a particular CA j was modelled separately (i.e. no information sharing) in an individual Bayesian analysis with minimally informative priors. This formulation can be thought of as the Bayesian equivalent of a "frequentist" analysis that does not take multiple testing into consideration. This should be comparable to performing a Fisher's exact test separately for each medication-CA combination and then using a P-value of 0.05 as the cut-off for statistical significance. However, as a Bayesian analysis places prior distributions on the model parameters, some shrinkage was expected due to the fact that the prior for the estimated PRR was centred on a PRR of 1 (no effect). This prior distribution will have some non-infinite variance, and the amount of shrinkage will also depend on the size of this variance. Estimates with low cell counts, for example, are more likely to be influenced by such a prior distribution as there is less information in the data itself for such combinations. These individual Bayesian models were therefore expected to identify less signals than the equivalent Frequentist models.

Table 6.2. Layout of cell counts c_{ij} for each medication-CA combination in a two-dimensional model for EUROmediCAT data with no information sharing.

		Congenital anomalies				
	CA (j) *ATC5* (i)	**1**	**2**	**…**	$\boldsymbol{n_j}$	***Total***
ATC medications	**1**	c_{11}	c_{12}	…	c_{1n_j}	$\boldsymbol{c_{1.}}$
	2	c_{21}	c_{22}	…	c_{2n_j}	$\boldsymbol{c_{2.}}$
	⋮	⋮	⋮		⋮	⋮
	$\boldsymbol{n_i}$	c_{n_i1}	c_{n_i2}	…	$c_{n_in_j}$	$\boldsymbol{c_{n_i.}}$
	Total	$\boldsymbol{c_{.1}}$	$\boldsymbol{c_{.2}}$	**…**	$\boldsymbol{c_{.n_j}}$	$\boldsymbol{N}$ $(= c_{..})$

Discrete groupings of either medications or congenital anomalies

The structure of the data when considering information sharing for medications using discrete groupings of four digit ATC3 codes is displayed in Table 6.3. In this setting, the effects for each group of medications were averaged across all the CAs, such that the CAs were treated as coming from one overall group (i.e. allowing a distribution of effects across the group of all CAs separately for each group of ATC3 medications). In Table 6.3, each d represents a group of medications according to their ATC3 codes. There were $d = 1, \dots, D$ groups, and within each group d, there were $i = 1, \dots, n_d$ unique ATC5 medication codes. In the other dimension were the $j = 1, \dots, n_j$ CAs as one group. A combination of a particular medication di with a particular CA j was denoted c_{dij}. Each group of medications d can be considered a set; for example, the grey shading in Table 6.3 represents the $d = 2$ group of medications across the n_j CAs.

Table 6.3. Example of data structure for a model of information sharing by discrete groupings of medications.

			Congenital anomalies					
		CA (j)	1	2	…	…	…	n_j
	ATC3 (d)	*ATC5 (i)*						
ATC medications	1	1						
		2						
		⋮						
		n_1						
	2	1						
		2						
		⋮						
		n_2						
	⋮	⋮						
	D	1						
		2						
		⋮						
		n_D						

The structure of the data when considering information sharing in the other dimension, i.e. for CAs only, is displayed in Table 6.4. In this model, the CAs were discretely grouped by their EUROCAT organ system class groupings, and the effects for each group of CAs were averaged across the medications (allowing a distribution of effects across the group of all medications separately for each group of CAs). Then there were A groups of CAs and $j = 1, \dots, n_a$ CAs within each group. In the other dimension were the $i = 1, \dots, n_i$ medications as one group. Any combination of a CA aj with a particular

medication i was denoted c_{iaj}. Here, each group of CAs a is considered a set; for example, the grey shading in Table 6.4 represents the $a = 2$ group of CAs across the n_i medications.

Table 6.4. Example of data structure for a model of information sharing by discrete grouping of CAs.

		Congenital anomalies												
	Organ class (a)	1				2				…	A			
	ATC5 (i) CA (j)	1	2	…	n_1	1	2	…	n_2		1	2	…	n_A
ATC medications	1													
	2													
	⋮													
	n_i													

Two-dimensional discrete groupings of both medications and congenital anomalies

The structure of the data when considering discrete groupings by both medications and CAs is displayed in Table 6.5. In this model, sets of cells were created according to crossings of the two variables. Again d represented groupings of medications by ATC3 codes; there were $d = 1, \dots, D$ groups and $i = 1, \dots, n_d$ unique ATC5 medication codes within each group d. In the other dimension a denoted groupings of CAs by their EUROCAT organ system classes. There were A groups of CAs and $j = 1, \dots, n_a$ CAs within each group. Any two-dimensional crossing of a group of CAs a with a group of medications d was then considered a set; for example, the lighter grey shading in Table 6.5 represents a set that is the crossing of the $d = 2$ group of medications with the $a = 2$ group of CAs a. Any combination of a particular medication di with a particular CA aj was denoted c_{diaj}. The group $a = 2$ may denote, for example, the "nervous system" group of CAs, in which there are $n_2 = 6$ CAs. Similarly, the group $d = 2$ may denote the N03A "Antiepileptics" group of ATC3 coded medications, within which there are $n_2 = 13$ distinct ATC5 codes in the data. Suppose that the second CA in $a = 2$ is spina bifida, and the first medication code in $d = 2$ is N03AA01; the cell c_{2122} then denotes the cell of interest in the analysis of the combination of medication N03AA01 with the CA spina bifida; this is the darkest shaded cell in Table 6.5. Each set in the hierarchy in Table 6.5 (i.e. one set is the cells shaded in lighter grey) had a group distribution from which the elements of that set were drawn, such that each medication-CA combination within that two-way group shared a common prior distribution. There was also a prior distribution for the set of all top-level sets i.e. an average across all CAs and/or medications in the table.

Table 6.5. Example of data structure for a model of information sharing by discrete grouping of both medications and CAs.

		Congenital anomalies											
	Organ class (a)	1				2				…	A		
	CA (j)	1	2	…	n_1	1	2	…	n_2		1	2	…	n_A
ATC3 (d)	*ATC5* (i)												
ATC medications 1	1												
	2												
	⋮												
	n_1												
2	1												
	2												
	⋮												
	n_2												
⋮	⋮												
D	1												
	2												
	⋮												
	n_D												

Table 6.6 demonstrates one possible two-dimensional set in the EUROmediCAT data, using the previously mentioned example of the nervous system CAs in combination with the antiepileptic medications. It can be seen that most of the information in the set for this example comes from the most common nervous system CA spina bifida, and from the three most common N03A medications (N03AF01, N03AG01 and N03AX01). The majority of the cell counts in this set contain little or no information (a zero cell count is indicated by a "-").

6.2.2. Specification of Bayesian models for signal detection analyses

Bayesian models were applied assuming a Poisson distribution according to the four types of groupings described in the previous section. Models for this chapter were also initially defined assuming a negative binomial distribution in order to ascertain which model provided a better fit for these data (since a negative binomial model can allow for more flexible modelling of the variance and can therefore be useful if there are large departures from the equidispersion assumption; see Chapter 3 for previous discussion on this).

Table 6.6 presents all model formulae and notation, showing how the observed counts c_{ij} for each combination of a medication i and a CA j were modelled using a Poisson or negative binomial distribution. The code used to specify these models in JAGS is presented in Appendix C1.

Table 6.6. Exposure counts for an example set of an ATC3 medication group and a group of congenital anomalies in the two-dimensional discrete grouping of EUROmediCAT data.

		Nervous System CAs					
		Anencephaly	Arhinencephaly/ Holoprosencephaly	Encephalocele	Hydrocephalus	Microcephaly	Spina Bifida
N03A (Antiepileptics)	N03AA01	-	-	-	-	1	-
	N03AA02	-	-	-	1	1	1
	N03AA03	-	-	-	-	-	1
	N03AB02	-	-	-	1	2	-
	N03AD01	-	-	-	-	-	-
	N03AE01	-	-	-	-	1	1
	N03AF01	1	-	-	3	2	11
	N03AF02	-	-	-	-	-	2
	N03AG01	2	-	-	7	3	32
	N03AG04	-	-	-	1	-	-
	N03AX09	-	-	-	4	1	10
	N03AX11	-	-	-	-	1	1
	N03AX12	-	-	-	1	-	1
	N03AX14	-	-	-	-	-	2
	N03AX15	-	-	-	-	-	
	N03AX16	-	-	-	-	-	-

Table 6.7. Notation for Bayesian models applied to observed counts of 523 medications and 55 CAs in chapter 6.

Grouping	Distribution	Model definition	Prior distributions for model parameters	Hyper-prior distributions for prior parameters
No grouping: $i = 1, \dots, n_i$ medications $j = 1, \dots, n_j$ CAs	Poisson	$c_{ij}\|\lambda_{ij} \sim Poisson(e^{\lambda_{ij}} E_{ij})$	$\lambda_{ij} \sim Normal(\mu_{ij}, \tau_{ij})$	—
	Negative Binomial	$c_{ij}\|\lambda_{ij} \sim NegativeBinomial(p_{ij}, \ r_{ij})$ $p_{ij} = e^{\lambda_{ij}} E_{ij} = PRR_{ij} E_{ij}$	$\lambda_{ij} \sim Normal(\mu_{ij}, \tau_{ij})$ $r_{ij} \sim Uniform(b1_r, b2_r)$	—
Discrete grouping by medications only: $d = 1, \dots, D$ ATC3 groups $i = 1, \dots, n_d$ medications in each group $j = 1, \dots, n_j$ CAs	Poisson	$c_{dij}\|\lambda_{dij} \sim Poisson(e^{\lambda_{dij}} E_{dij})$	$\lambda_{dij} \sim Normal(\theta_{\lambda d}, t_{\lambda d})$ $t_{\lambda d} = 1/\sigma^2_{\lambda d}$	$\theta_{\lambda d} \sim Normal(\mu_\theta, \tau_\theta)$ $\sigma_{\lambda d} \sim Uniform(b1_\sigma, b2_\sigma)$
	Negative Binomial	$c_{dij}\|\lambda_{dij} \sim NegativeBinomial(p_{dij}, r)$ $p_{dij} = e^{\lambda_{dij}} E_{dij} = PRR_{dij} E_{dij}$	$\lambda_{dij} \sim Normal(\theta_{\lambda d}, t_{\lambda d})$ $t_{\lambda d} = 1/\sigma^2_{\lambda d}$ $r \sim Uniform(b1_r, b2_r)$	$\theta_{\lambda d} \sim Normal(\mu_\theta, \tau_\theta)$ $\sigma_{\lambda d} \sim Uniform(b1_\sigma, b2_\sigma)$
Discrete grouping by CAs only: $i = 1, \dots, n_i$ medications $a = 1, \dots, A$ groups of CAs $j = 1, \dots, n_a$ CAs in each group	Poisson	$c_{iaj}\|\lambda_{iaj} \sim Poisson(e^{\lambda_{iaj}} E_{iaj})$	$\lambda_{iaj} \sim Normal(\theta_{\lambda a}, t_{\lambda a})$ $t_{\lambda a} = 1/\sigma^2_{\lambda a}$	$\theta_{\lambda a} \sim Normal(\mu_\theta, \tau_\theta)$ $\sigma_{\lambda a} \sim Uniform(b1_\sigma, b2_\sigma)$
	Negative Binomial	$c_{iaj}\|\lambda_{iaj} \sim NegativeBinomial(p_{iaj}, r)$ $p_{iaj} = e^{\lambda_{iaj}} E_{iaj} = PRR_{iaj} E_{iaj}$	$\lambda_{iaj} \sim Normal(\theta_{\lambda a}, t_{\lambda a})$ $t_{\lambda a} = 1/\sigma^2_{\lambda a}$ $r \sim Uniform(b1_r, b2_r)$	$\theta_{\lambda a} \sim Normal(\mu_\theta, \tau_\theta)$ $\sigma_{\lambda a} \sim Uniform(b1_\sigma, b2_\sigma)$
Discrete grouping by medications and CAs: $d = 1, \dots, D$ ATC3 groups $i = 1, \dots, n_d$ medications in each group $a = 1, \dots, A$ groups of CAs $j = 1, \dots, n_a$ CAs in each group	Poisson	$c_{diaj}\|\lambda_{diaj} \sim Poisson(e^{\lambda_{diaj}} E_{diaj})$	$\lambda_{diaj} \sim Normal(\theta_{\lambda da}, t_{\lambda da})$ $t_{\lambda da} = 1/\sigma^2_{\lambda da}$	$\theta_{\lambda da} \sim Normal(\mu_\theta, \tau_\theta)$ $\sigma_{\lambda da} \sim Uniform(b1_\sigma, b2_\sigma)$
	Negative Binomial	$c_{diaj}\|\lambda_{diaj} \sim NegativeBinomial(p_{diaj}, r)$ $p_{diaj} = e^{\lambda_{diaj}} E_{diaj} = log(PRR_{diaj})\ E_{diaj}$	$\lambda_{diaj} \sim Normal(\theta_{\lambda da}, t_{\lambda da})$ $t_{\lambda da} = 1/\sigma^2_{\lambda da}$ $r \sim Uniform(b1_r, b2_r)$	$\theta_{\lambda da} \sim Normal(\mu_\theta, \tau_\theta)$ $\sigma_{\lambda da} \sim Uniform(b1_\sigma, b2_\sigma)$

Description of models in Table 6.7

Each form of λ according to the different groupings presented in Table 6.6 denotes the $log(PRR)$ for the combination of a medication i and a CA j. The negative binomial model essentially introduces an additional "dispersion parameter", which can be used to adjust the variance independently of the mean. There are a number of different characterisations and formulations of the negative binomial distribution, which have been discussed in detail by Cameron and Trivedi [2013]. For these analyses, a Poisson-gamma mixture was used to specify the negative binomial model. This model is based on the assumption that the data follow a Poisson distribution with additional Gamma-distributed unobserved individual heterogeneity, which is described by an additional dispersion parameter r (this is also known as the shape parameter, see Table 6.6). As $r \to \infty$, the negative binomial distribution tends to a Poisson distribution, where if $k = \frac{1}{r}$ then $k = 0$ represents no overdispersion. The parameters of the negative binomial distribution p and r were given prior distributions through λ.

Prior distributions for model parameters

For individual Poisson and negative binomial models, each λ followed a Normal prior distribution with mean μ and precision τ. A choice of $\mu = 0$ represents a distribution centred on an average $log(PRR)$ of zero, i.e. an average PRR of 1, or of no effect. Suppose that most of the effects are believed likely to have a PRR of between $\frac{1}{30}$ and 30, corresponding to a $log(PRR)$ of between -3.4 and 3.4. This choice of limits for the prior variation in the PRRs was somewhat arbitrary; however it was considered unlikely that effects would be larger than a PRR of 30 or smaller than $\frac{1}{30}$ (particularly given that only a small number of protective associations are expected; see the following section for further discussion on this). The standard error for each λ was then $\sigma = \frac{3.4}{1.96} = 1.74$, with the corresponding choice of $\tau = \frac{1}{\sigma^2} = \frac{1}{1.74^2} = 0.33$ therefore being used as the parameter value for the precision of the prior distribution for λ. Other values of τ were also considered in order to assess the sensitivity of models to these choices of prior parameters. For negative binomial models, the limits of the uniform prior for each overdispersion parameter r were chosen to ensure that this was always positive, i.e. $b1_r = 0$ and with a large value for the upper limit such as $b2_r = 1000$. A relatively large value for $b2_r$ was chosen to ensure that r was allowed to tend to ∞ in each type of model, and therefore converge to a Poisson

distribution if that was a better fit. Different choices of $b2_r$ were assessed for their effect on model fit and results.

Hyper-prior distributions for prior parameters

For BHMs grouping medications and CAs, the $log(PRR)$ for each medication-CA combination followed a Normal prior distribution mean θ, representing the average $log(PRR)$ across the group to which each medication-CA belonged. Similarly, σ^2 denoted the variance of the average $log(PRR)$ the group to which each medication-CA belongs (and with t the related precision), i.e. representing the variation in estimates of combinations within the same group. These two parameters were then given their own prior distributions, using a Normal distribution for θ and a uniform distribution for σ^2. Normal distributions for θ were given minimally informative hyper-parameters centred on zero and with relatively large variance, i.e. $\mu_{\theta,} = 0$ and $\tau_\theta = 0.33$ (as described previously). A uniform prior distribution is commonly used for the standard deviation of variance parameters in BHMs [Gelman A., 2006]; the limits of the uniform distribution used for σ ensured that the variance was positive whilst also allowing σ to take a relatively large value, i.e. $b1_\sigma = 0$ and $b2_\sigma = 100$. Different choices of $b2_\sigma$ were also assessed for their effect on model fit and results.

6.2.3. Further characteristics of Bayesian models for signal detection

Choice of parameters for prior distributions

The amount of shrinkage in a hierarchical model can depend on the choice of parameters for the prior distributions. Minimally informative priors were therefore used for all parameters and hyper-parameters, as described above, in order to assess the effect of the groupings themselves on the model results, as opposed to measuring the effect of the choice of prior distribution. The limits used for the prior distribution of the $log(PRR)$ were based on the assumption that the PRRs are not likely to be smaller than $\frac{1}{30}$ or greater than 30. One of the strongest known teratogenic associations is that of valproic acid and spina bifida; the increase in risk associated with taking valproic acid in first trimester of pregnancy has been estimated to be around 13 times higher compared to the risk for women that took no antiepileptic medications, with a 95% CI ranging from around 8 up to 21 times an increase in risk [Jentink et al., 2010]. It was therefore considered unlikely that any signal would involve a risk much larger than this upper 95% CI value of 21, hence the selected value of 30 was thought to be a sensible upper limit for the PRR in this chapter. As a quick check to see whether this assumption seemed reasonable for this data in practice,

approximate 95% confidence limits for the observed PRR for each combination with at least 3 exposures were calculated as follows (see Table 4.1 for notation)

$$PRR_{ij} = \frac{c_{ij} \,/\, c_{i.}}{c_{i'j} \,/\, c_{i'.}}$$

$$SE(logPRR) = \sqrt{\frac{1}{c_{ij}} - \frac{1}{c_{i.}} + \frac{1}{c_{i'j}} - \frac{1}{c_{i'.}}}$$

$$95\%\ CI = e^{\log(PRR_{ij}) \pm 1.96\ x\ SE(log(PRR_{ij}))}$$

Using this approximation, none of the observed PRRs had a lower 95% confidence limit lower than 1/30, whilst only seven (all with a count of only 3 exposures) had an upper 95% confidence limit greater than 30. The assumption that the PRRs are not likely to be smaller than $\frac{1}{30}$ or greater than 30 was therefore considered reasonable for use in prior parameters in BHMs for this chapter.

Overdispersion in count models

An important assumption of the Poisson distribution for these data is that the conditional mean and variance should be equal, i.e. that the data are equi-dispersed. One way in which this assumption might be violated is if there is an underlying structure to the data that leads to a lack of independence in the exposure counts, these being medications and the number of specific types of CAs; a woman may have more than one count in the database for either or both of these. It is often the case that a malformed foetus has more than one major CA, and such co-occurrences are not necessarily independent. Respiratory and ear, face and neck defects, for example, were found to be the types of CA most likely to occur in combination with other CAs [Calzolari et al., 2014]. Of the 5.2 million births in the EU each year, around 2% are babies with at least one CA; if multiple CAs in the same baby occurred independently of each other, then the risk of a second CA in an affected pregnancy should be approximately 2% of those with at least one CA. In the data for this thesis, around 300 (2%) of the 15,058 malformed foetuses would then be expected to have a second CA; the expected number of foetuses with 3 CAs would be around 6 (0.04%), and less than one foetus (0.0008%) would be expected to have 4 or more CAs. However, Table 4.6 showed that these numbers were actually much higher in the data, with 1,431 (9.5%), 339 (2.4%) and 117 (0.8%) of the 15,058 malformed foetuses having 2, 3 and 4 or more CAs, respectively. Under the assumption of independence, the observed number of multiple CAs would therefore exceed the expected number by a considerable amount. The use of

multiple medications during pregnancy has also become increasingly common. A study in the US, for example, found that that the average number of medications used in the first trimester of pregnancy increased from 1.6 in 1976-1978 to 2.6 in 2006-2008, with 82.3% of women in 2008 taking at least one medication during this stage of their pregnancy [Mitchell et al., 2011]. Combinations of medications taken together may not be independent; in the management of asthma, for example, inhaled short-acting beta-2-agonists are taken for symptom relief and these are often used alongside inhaled corticosteroids and other medications [Garne et al., 2016]. As the structure of the observed data may therefore violate the independence assumptions for both the CA and medication exposure counts in the BHMs used for this chapter, an alternative way of modelling the exposure counts was also considered by using the number of exposed individuals as the marginal totals, instead of the sum of the exposure counts.

Definition of signals in Bayesian models

Signals were identified using the 2.5th percentile values of the posterior distribution thresholds as a potential cut-off, representing the lower limit of the 95% PCIs. In this way, any medication-CA combination for which the posterior 2.5th percentile value for the $log(PRR)$ was greater than zero was considered a signal. This represents the situation where 97.5% of the posterior distribution for the $log(PRR)$ of interest lies above zero, or equivalently that 97.5% of the posterior distribution for the PRR of interest lies above one. Equally, if a 95% PCI for a $log(PRR)$ includes the value of zero, this would not be inconsistent with there being no effect of that particular medication on that particular CA (i.e. a PRR of 1). As with the frequently used 5% level of significance in frequentist hypothesis tests, this choice of cut-off value is arbitrary. The effect of choosing a stricter 0.5th percentile as a cut-off (corresponding to the lower limit of a 99% PCI) on the resulting set of signals for each model was also assessed. Note that by using the lower limit of the posterior distribution, any "protective" associations that may arise were thus disregarded, as seen in the previous chapter with the use of a one-sided Fisher's exact test. Whilst such associations were not identified as potential signals (since the point of these analyses is to identify teratogens), the number of such associations were again monitored to check that they were not occurring more frequently than expected. This was done by identifying those combinations with a 97.5th (or 99.5th) percentile value for the posterior distribution of the $log(PRR) < 0$, again corresponding to the 95% (or 99%) PCIs.

6.2.4. Summary of models applied to EUROmediCAT data in this chapter

In summary, the following models were applied to the EUROmediCAT data, each using a Bayesian analysis with minimally informative priors

1. No information sharing (individual Bayesian analyses)
2. Discrete information sharing for medications
3. Discrete information sharing for CAs
4. Discrete information sharing for both medications and CAs

Poisson and negative binomial models were first compared, in order to ascertain which model offered an improved fit to the structure of the data. All models as specified in Table 6.6 were run using JAGS via the programme R and its package `rjags`. As described previously, the `coda` package [Plummer Martyn et al., 2003] was used to assess model convergence and to summarise the sample posterior distribution for each parameter. This included calculation of convergence statistics as well as visual inspection of trace, density and auto-correlation plots for the parameters in each model. These measures were also used to determine the required number of total iterations and thinning. See Chapter 3 (section 3.5.2) for further details on the assessment of convergence for Bayesian models.

As described in chapter 4, results from all models were compared in terms of their identification rate and "high risk" proportion according to the Australian risk categorisation system for ATC coded medications. Briefly, these measures were defined as

$$\textbf{Identification rate} = \frac{\text{Number of "high risk" medications identified as signals}}{\text{Total number of "high risk" medications in the data}}$$

$$\text{"}\textbf{High risk}\text{" }\textbf{proportion} = \frac{\text{Number of medication signals in "high risk" category}}{\text{Total number of medications identified as signals}}$$

$$\textbf{Effective workload} = \text{Total number of medication signals}$$

Results from Bayesian models were also compared with those obtained using the single and double FDR methods with groupings by ATC3 medications and an FDR cut-off of 50% (see Chapter 5).

6.3. Results

Chapter 4 (section 4.4.2) described and summarised the EUROmediCAT dataset used for analyses in this thesis. Briefly, data on 15,058 foetuses was available, with 55 CAs and 523 ATC medications being monitored for signal detection purposes. This included 26,765 total exposure counts across 28,765 possible medication-CA combinations.

6.3.1. Assessment of Bayesian hierarchical models

Convergence of parameters in BHMs was assessed by visual inspection of autocorrelation, trace and density plots for all models (data not shown). For each type of grouping and model, an adaptive phase of 1000 iterations was used, followed by a burn-in of 1000 iterations being discarded. Models were then run using three chains each with 20,000 iterations. A large thin was used due to high levels of autocorrelation, meaning that only one in every 20 successive iterations was used to summarise the posterior distribution. The thinned posterior samples generally demonstrated good mixing of chains and low autocorrelation. Posterior density distributions were generally reasonably symmetric, although some distributions were skewed.

6.3.2. Modelling count data using Bayesian hierarchical models

Estimating the presence of dispersion using a negative binomial model

Table 6.8 summarises the posterior distributions for the estimated dispersion parameter r for the negative binomial models according to different choices of parameters for the prior distribution for r.

Table 6.8. Posterior distribution of dispersion parameter r in negative binomial models grouped by ATC3 medications and/or congenital anomaly (CA) subgroups.

Prior parameters for r	**Type of grouping in model**	**Median of posterior distribution for r (95% PCI)**
Uniform(0, 100)	CAs	71 (39 – 98)
	Medications	89 (63 – 100)
	Medications and CAs	94 (76 – 100)
Uniform(0, 1000)	CAs	500 (81 – 976)
	Medications	633 (167 – 982)
	Medications and CAs	773 (336 – 990)
Uniform(0, 10000)	CAs	5124 (326 – 9817)
	Medications	5385 (690 – 9802)
	Medications and CAs	5891 (916 – 9766)

Across these different scenarios, the parameter r had a distribution that generally approached the chosen upper limit, indicating that r tended to a large value (i.e. $\frac{1}{r}$ was

close to zero), in which the negative binomial distribution should be approximately similar to a Poisson distribution. It is therefore unlikely that the dispersion parameter substantially improves the model fit for these data, implying that the use of a Poisson model was reasonable for these data. This was checked by confirming that the use of a Poisson rather than a negative binomial model did not have any important effect on the actual model estimates (i.e. the estimated PRRs). To do this, Poisson and negative binomial models were both implemented for each type of grouping in order to assess which was a better fit for the data, and whether there was any evidence of overdispersion. Results are presented in the following section.

Comparison of estimates from negative binomial and Poisson models

The "high risk" proportion is plotted against the effective workload for Poisson (filled grey symbols) and negative binomial (hollow black symbols) models in Figure 6.1, using a 95% PCI (panel A) and a 99% PCI (panel B) as cut-offs for defining signals. Both axes are plotted on a logarithmic scale, such that the distance between two consecutive horizontal gridlines represents a doubling in "high risk" proportion. Note that because the "high risk" proportion is calculated as a percentage of the effective workload, this is sensitive to small changes at the lower scale of the x-axis (i.e. the left hand side of Figure 6.1). Differences between points towards the lower end of the scale on both axes therefore look greater than differences of the same size at the top of the scales. For example, grouping by type of CA using a negative binomial model identified four fewer medication signals than a Poisson model with the same grouping (4 compared to 8) when using a 95% PCI cut-off to classify associations as signals (Figure 6.1 panel A). For models without any grouping the two markers appear much closer together, but the absolute difference in effective workload between the negative binomial and Poisson model for this grouping is in fact only six (total of 153 compared to 159 medication signals). Although this looks far greater on the scale used here, this is similar to the difference between the negative binomial and Poisson models using CA groupings. The choice of a 99% PCI cut-off for definition of signals shown in Figure 6.1 shows that smaller numbers of medication signals are identified by this stricter threshold, but with better "high risk" proportions. This can be seen in that all points in panel A of Figure 6.1 are higher and further left than for the same models and groupings shown in panel B. When grouping by CAs and using a 99% PCI, the negative binomial and Poisson models both gave three medication signals (hence the markers for these models are in the exact same point on Figure 6.1). Of these three medications, one was "high risk", i.e. the "high risk" proportion was 33%.

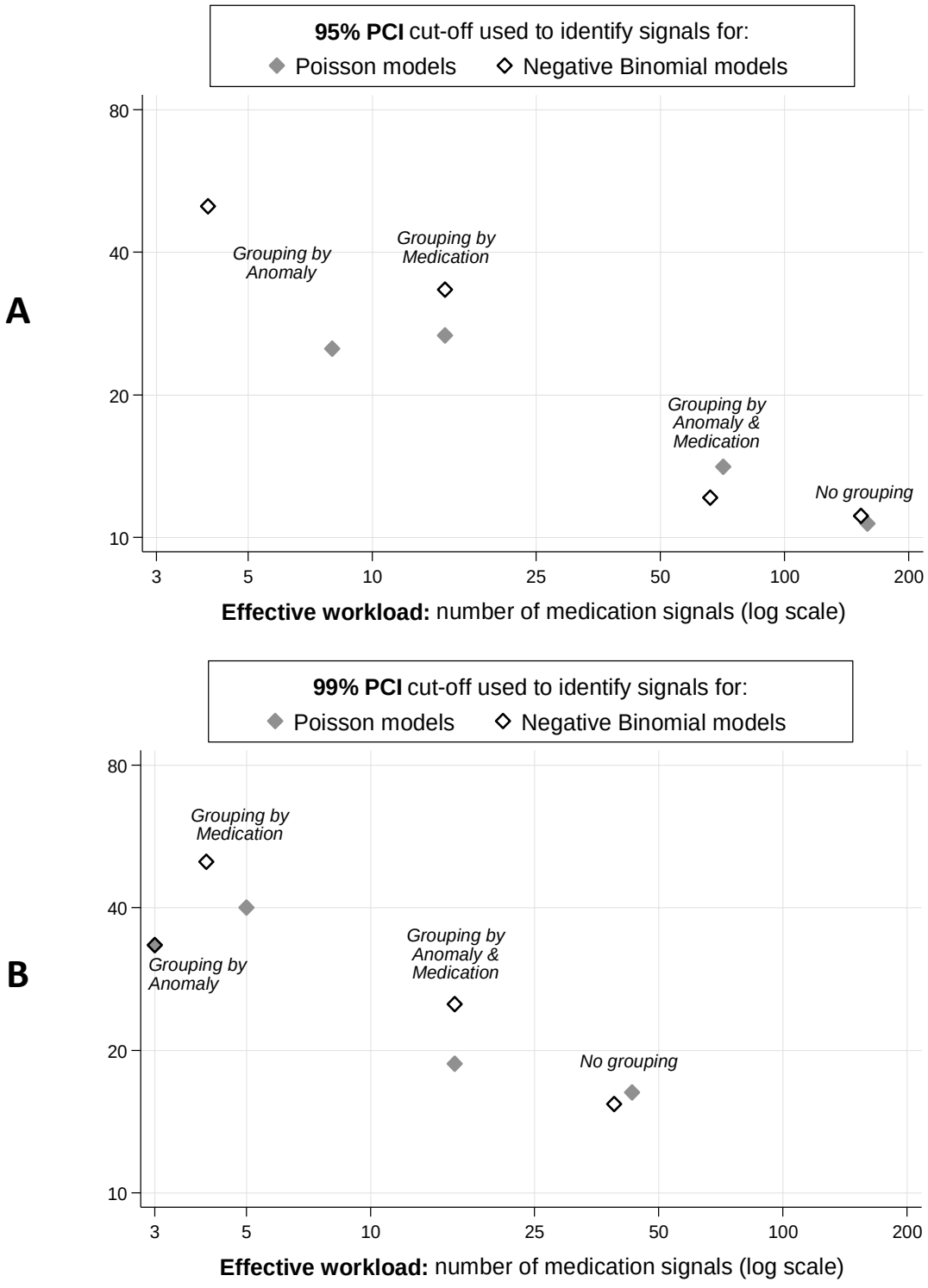
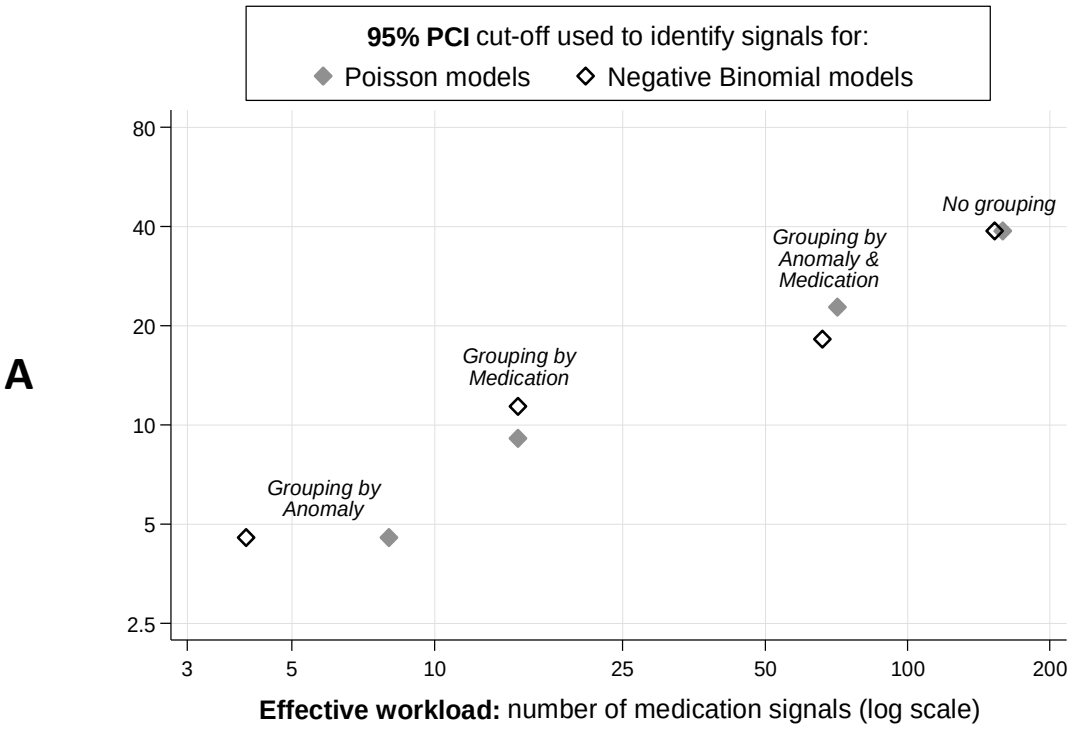


Figure 6.1. "High risk" proportion (percent of all medication signals that are in the "high risk" category) vs. effective workload comparing the use of Poisson (filled grey markers) and negative binomial (hollow black markers) distributions to model the cell counts, using (A) 95% PCI and (B) 99% PCI to define signals in Bayesian models for four types of grouping.

For some groupings, the negative binomial resulted in improved “high risk” proportion compared to the Poisson model, but these differences were not consistent across the groupings or cut-offs (Figure 6.1). All differences in “high risk” proportion were due to only one or two additional “high risk” signals being identified by the negative binomial model, or to the same number of “high risk” signals but a greater number of medication signals overall affecting the “high risk” proportion. The differences in “high risk” proportion and effective workload between the negative binomial and Poisson models were small relative to the effect of the type of grouping used by either model.

The Identification rate is plotted against the effective workload in Figure 6.2 for all Poisson (filled grey symbols) and negative binomial (hollow black symbols) models, using a 95% PCI (panel A) and a 99% PCI (panel B) as cut-offs to define signals. As with the “high risk” proportion, the identification rate was similar for comparable effective workloads when using a Poisson or a negative binomial model. Again, the 99% PCI cut-off for definition of signals shown in panel B of Figure 6.2 shows that smaller numbers of medication signals are given by this stricter threshold, and with lower proportions of the total potential “high risk” medications in the dataset being identified as signals (when compared to the use of a 95% PCI to define the signals, as shown in panel A).

In summary, Figure 6.1 and Figure 6.2 demonstrate that there were no substantial differences in results whether using a negative binomial or Poisson distribution to model the count data. Instead, the main differences in the metrics presented arise from the type of grouping and the choice of cut-off level of PCI for assigning combinations as signals, and these are explored in detail in the following sections. Furthermore, the Poisson model is more parsimonious (has less parameters) and straightforward to implement compared with the negative binomial models. For the rest of this chapter, therefore, results are presented for analyses using a Poisson distribution to model the observed counts.



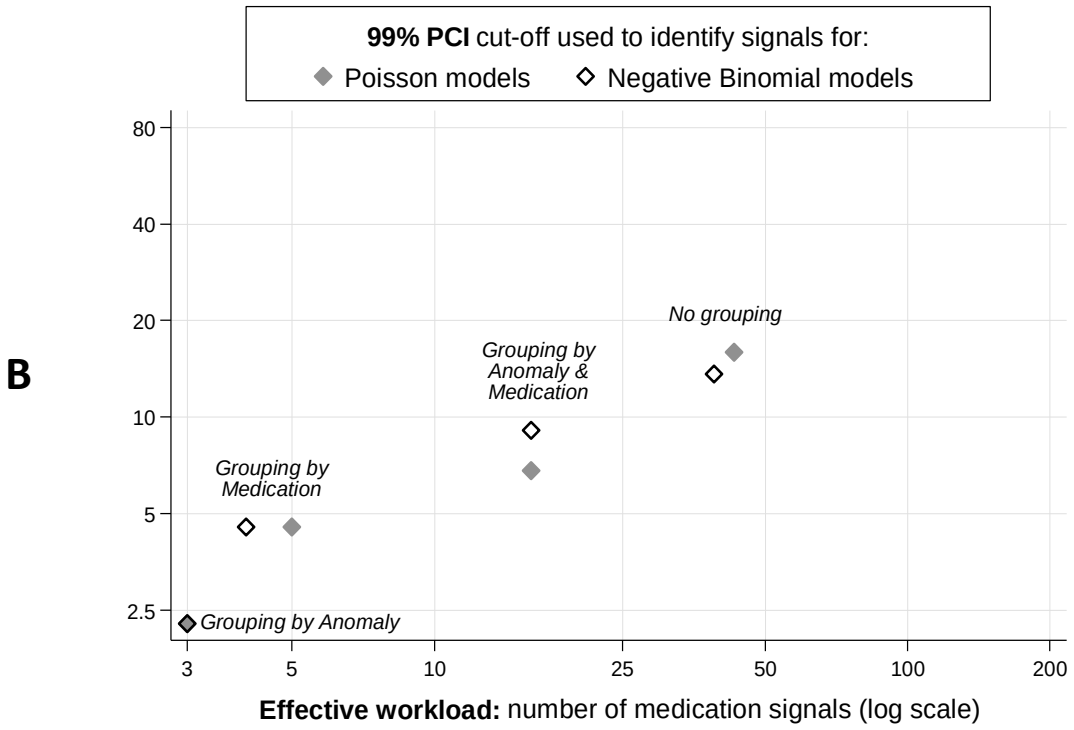


Figure 6.2. Identification rate (proportion of all "high risk" medications that are identified as signals) vs. effective workload comparing the use of Poisson (filled grey markers) and negative binomial (hollow black markers) distributions to model the cell counts, using (A) 95% PCI and (B) 99% PCI to define signals in Bayesian models for four types of grouping.

Calculation of expected counts

As a sensitivity analysis, two variants on the calculation of the expected count used to determine the expected-to-observed ratio (i.e. the PRR) were assessed. Firstly, when considering the potential violation of the assumption of independence of exposure counts in the data, the number of individuals were used as the marginal totals for calculation of the E_{ij}, instead of the total number of exposures. In practice this meant that a malformed foetus with two CAs and an exposure to one medication, for example, would only be counted once in the marginal (row) total for that medication, despite being present in two separate CAs (once in each of two columns). Similarly, a malformed foetus with two medication exposures and one CA would only be counted once in the marginal total for that CA, despite being counted in each row for the two medications. One problem with this approach was that the sum of the marginal totals for the medications and the sum of the marginal totals for the CAs do not add up, therefore the choice of the overall total N (see Table 4.1) was not clear. A number of different values were therefore used for N to see if this approach was feasible, including a simple average of the two sums of marginal totals, an average weighted by the total number of medications and CAs and using just one or the other sum of the marginal counts. For all choices of N considered, however, it was found that the expected counts were always much lower than the observed counts. Although the number of "protective" associations did completely diminish when using this alternative calculation for the expected count, the total number of signals increased hugely to the point where BHMs produced thousands of signals (data not shown). This did not represent a realistic resulting workload, and with such a high number of potential signals a very high number of false positives would be likely. A second alternative definition of the expected count was also considered to address the fact that the exposure count for each particular medication-CA combination was included twice in the calculation of each E_{ij} (see section 6.2 for details). Use of this alternative calculation for E_{ij} did not, however, have a material effect on the results for any of the BHMs; although one or two additional "high risk" medications were identified for a number of the BHMs when using this alternative definition, this was at the expense of an increased total number of signals for very similar "high risk" proportion and identification rates (data not shown). Since neither of these variants of the expected count appeared to improve the BHMs used, all results in this chapter are presented for models which used the original calculation $E_{ij} = \frac{c_{i.}\, c_{.j}}{N}$.

6.3.3. Signal detection using Poisson Bayesian hierarchical models, with comparison to single and double FDR procedures

A summary of the main results from Fisher's exact test (see Chapter 5) and Poisson BHMs are displayed in Table 6.9. Different cut-offs used to define signals are displayed, starting with more "lenient" choices (identifying a greater total number of signals) for each method; the first method, for example, shows results from Fisher's exact test where combinations were defined as signals if they had an unadjusted P-value <0.05 (first row of Table 6.9) or <0.01 (second row). For FDR methods, the different cut-off levels correspond to the 5 points for each of these methods in Figure 6.3 and Figure 6.4. FDR methods used cut-offs ranging from 5% to 50% in 5% increments; for some cut-off levels, however, it took more than one 5% increment to see a change in the resulting number of signals. For example, the set of signals using single FDR was the same for any FDR cut-off value between 15-25%. This is why there are only 5 (rather than 10) points for each method in Table 6.9, Figure 6.3 and Figure 6.4. Table 6.9 shows the number of ATC3 groups without any signals (column "ATC3 groups with no signals") increased as the cut-off level for each method became stricter and fewer signals were identified (column "Combinations identified as signals: total"). Fisher's test and individual BHMs (i.e. no grouping or adjustment for multiple testing) identified the most signals across more groups. The number of medication signals is also shown for each method (effective workload, column in bold), including a breakdown by the type of risk category according to the Australian classification system. More "high risk" DX medications were identified by use of individual Poisson BHMs; this method also, however, identified the greatest overall number of combinations signals and gave an effective workload of 159 unique medications to follow up, representing 30% of all medications in the analysis. The "strictest" method was single FDR with a cut-off of 5%, identifying only two medication signals, of which one was "high risk". The "high risk" proportion, identification rate and effective workload were similar using (i) a BHM grouped by both medications and CAs with a 95% PCI cut-off, and (ii) Fisher's exact test with an unadjusted P-value cut-off P<0.01. The only models with higher identification rates were individual BHMs with a 95% PCI cut-off and Fisher's exact test with an unadjusted P-value cut-off P<0.05; both these methods identified an extra 10% of the "high risk" signals, but at the expense of a two-fold increase in effective workload (compared to methods with the next highest workload). Compared to double FDR with a cut-off of 50%, individual BHMs with 95% PCI cut-offs and unadjusted P-value<0.05 models both gave an effective workload almost 10 times higher for a gain in identification rate of only 10% (around a 1.5 fold increase).

Table 6.9. Summary of results from Fisher's exact test with various adjustments for multiple testing, and from BHMs with a Poisson distribution.

Type of model and grouping	Cut-off level used to define signals	ATC3 groups with no signals [a]	**Combinations identified as signals**		**Number of unique medication signals**				"High risk" proportion (%) [c]	Identification rate (%) [d]	CAs with at least one signal [e]
			Number with < 3 exposures	Total	"Low risk" category: A, B or C	"High risk" category: D or X	No risk category	**Total** [b]			
Fisher's test: **no grouping or adjustment** for multiple testing	P <0.05	56	354 [f]	252	100	16	39	**155**	10	36	49
	P <0.01	79	54 [f]	91	38	11	18	**67**	16	25	34
Fisher's test: **single FDR (no grouping)**	FDR 50% [g]	109	0	10	3	3	2	**8**	38	7	9
	FDR 30%	110	0	9	2	3	2	**7**	43	7	8
	FDR 20% [g]	111	0	7	1	3	2	**6**	50	7	6
	FDR 10%	112	0	4	1	1	2	**4**	25	2	4
	FDR 5%	114	0	2	0	1	1	**2**	50	2	2
Fisher's test: **double FDR** with **ATC3 grouping**	FDR 50% [g]	109	1	25	5	6	5	**16**	38	14	14
	FDR 30% [g]	109	0	19	4	6	5	**15**	40	14	11
	FDR 20% [g]	112	0	13	3	5	3	**11**	45	11	8
	FDR 10%	114	0	7	0	4	1	**5**	80	9	5
	FDR 5%	114	0	5	0	2	1	**3**	67	5	4

[a] out of a total of 116 ATC3 groups of medications

[b] out of a total of 523 unique ATC coded medications

[c] "High risk" proportion: the proportion of all medication signals that are "high risk" category D or X medications

[d] Identification rate: the proportion of "high risk" category D or X medications that are identified as signals, out of n=44 category D or X medications in the dataset

[e] out of a total of 55 CAs

[f] Combinations with less than 3 exposures are not considered signals when using Fisher's exact test

[g] FDR cut-off levels were assessed in 5% increments from 5% to 50%, but some cut-off levels resulted in the same values for this table; for single FDR cut-offs in the ranges 15-25% and 35-50%, and for double FDR 15-25%, 30-40% and 45-50% all provided the same number of signals

Table 6.9 (continued). Summary of results from Fisher's exact test with various adjustment for multiple testing, and from BHMs with a Poisson distribution.

Type of model and grouping	Cut-off level used to define signals	ATC3 groups with no signals [a]	**Combinations identified as signals**		**Number of unique medication signals**				"High risk" proportion (%) [c]	Identification rate (%) [d]	CAs with at least one signal [e]
			Number with < 3 exposures	Total	"Low risk" category: A, B or C	"High risk" category: D or X	No risk category	**Total** [b]			
Individual Poisson BHMs: **no grouping**	95% PCI	47	59	223	98	17	44	**159**	11	39	48
	99% PCI	88	2	53	18	7	18	**43**	16	16	24
Poisson BHM: discrete grouping by **ATC3**	95% PCI	107	0	21	3	4	8	**15**	27	9	9
	99% PCI	112	0	7	1	2	2	**5**	40	5	6
Poisson BHM: discrete grouping by **CA**	95% PCI	110	0	10	2	2	4	**8**	25	5	6
	99% PCI	113	0	3	1	1	1	**3**	33	2	3
Poisson BHM: discrete grouping by **ATC3 & CA**	95% PCI	70	44	112	33	10	28	**71**	14	23	36
	99% PCI	105	1	24	6	3	7	**16**	19	7	14

[a] *out of a total of 116 ATC3 groups of medications*

[b] *out of a total of 523 unique ATC coded medications*

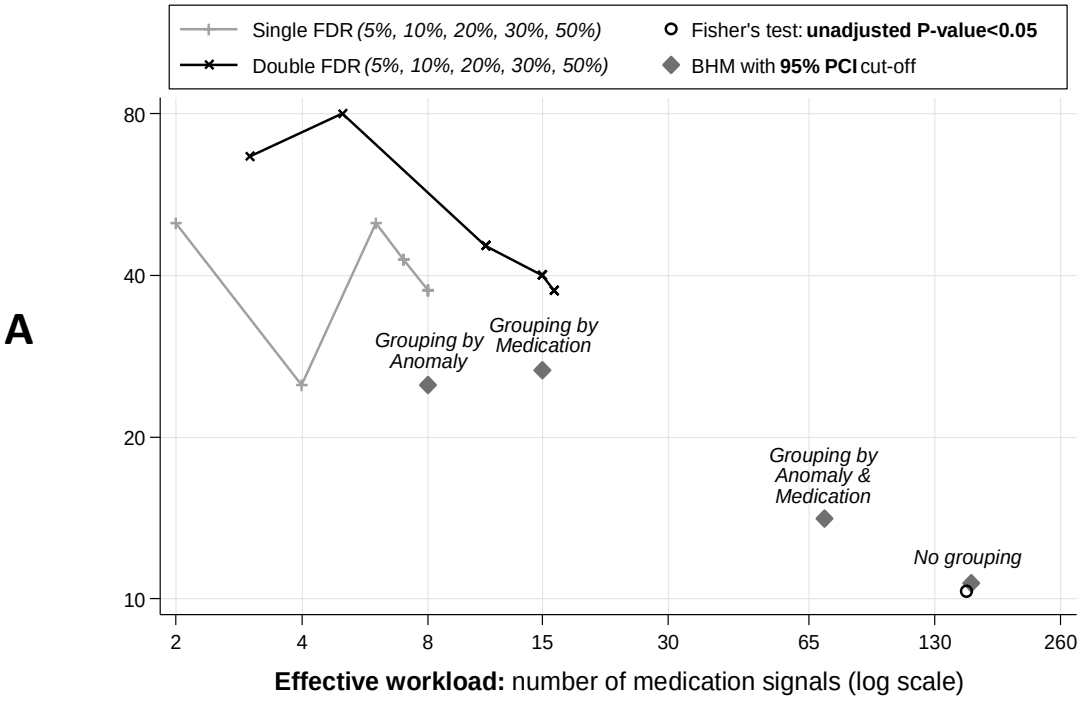
[c] *"High risk" proportion: the proportion of all medication signals that are "high risk" category D or X medications*

[d] *Identification rate: the proportion of "high risk" category D or X medications that are identified as signals, out of n=44 category D or X medications in the dataset*

[e] *out of a total of 55 CAs*

The “high risk” proportion for all models displayed in Table 6.9 is plotted against the effective workload in Figure 6.3 using a 95% PCI (panel A) and a 99% PCI (panel B) as a cut-off to define signals, with results for FDR cut-offs of 5-50% for both FDR methods of P-value adjustment as previously described. Similarly, the identification rate is plotted against the effective workload in Figure 6.4, again using a 95% PCI (panel A) and a 99% PCI (panel B) to define signals. These figures show that the double FDR resulted in the highest “high risk” proportion and identification rate for comparable effective workloads when using other methods. Whilst the ungrouped BHMs and those grouped discretely by both medications and CAs sometimes identified greater numbers of “high risk” medications, this was at the expense of a substantial increase in effective workload.

Figure 6.5 and Figure 6.6 are heatmaps showing which combinations of CAs and ATC2 medication groups involved signals from single and double FDR, respectively, each using an FDR cut-off of 50%. Figure 6.7, Figure 6.8 and Figure 6.9 show the same information for BHMs using a 95% PCI cut-off, with groupings by CAs, by medications and by both medications and CAs, respectively. In Figure 6.5 to Figure 6.9, the shading represents the number of medication-CA combinations that are a signal in each grouping of and ATC2 medication with a CA, according to each method. The N03 and G03 medication groups contained the most signals for all methods, followed by the A10 group of medications, which included signals in all methods except for the BHM grouped by type of CA.



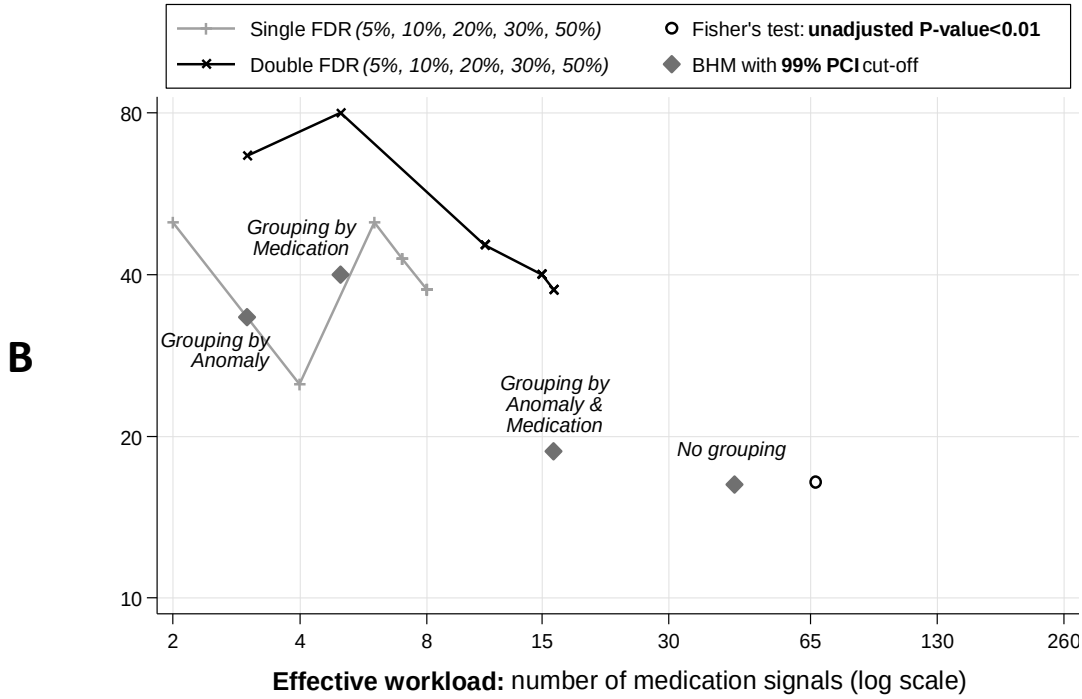
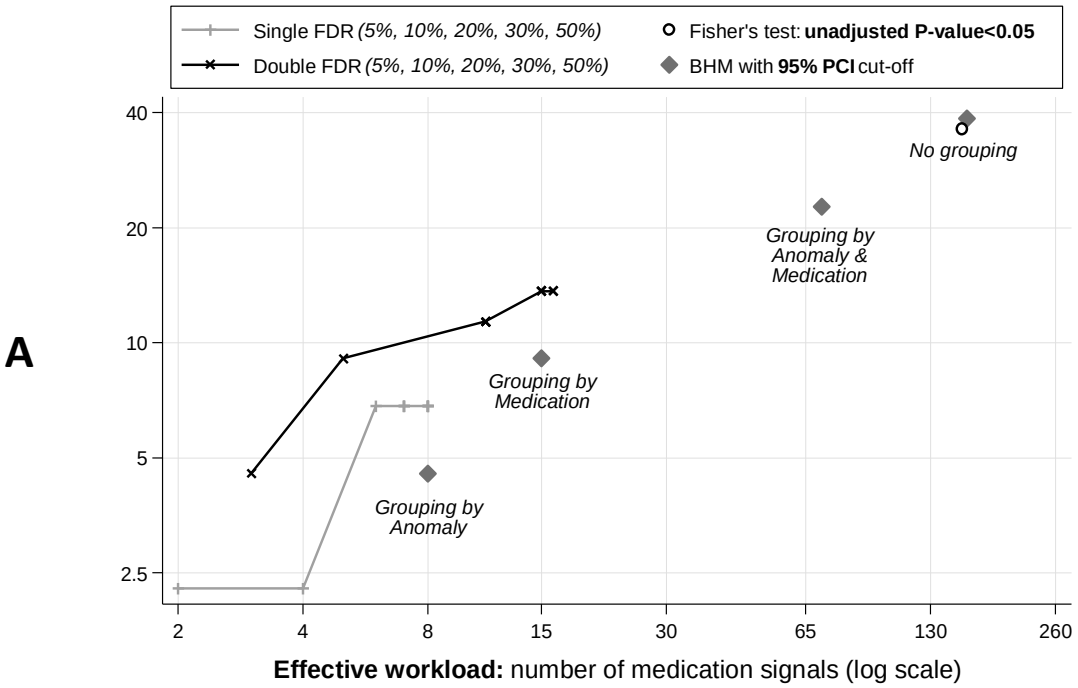


Figure 6.3. "High risk" proportion (percent of all medication signals that are in the "high risk" category) vs. effective workload: comparing the use of single and double FDR procedures with Poisson BHMs using (A) 95% PCI and (B) 99% PCI as a cut off for definition of signals for four types of grouping.



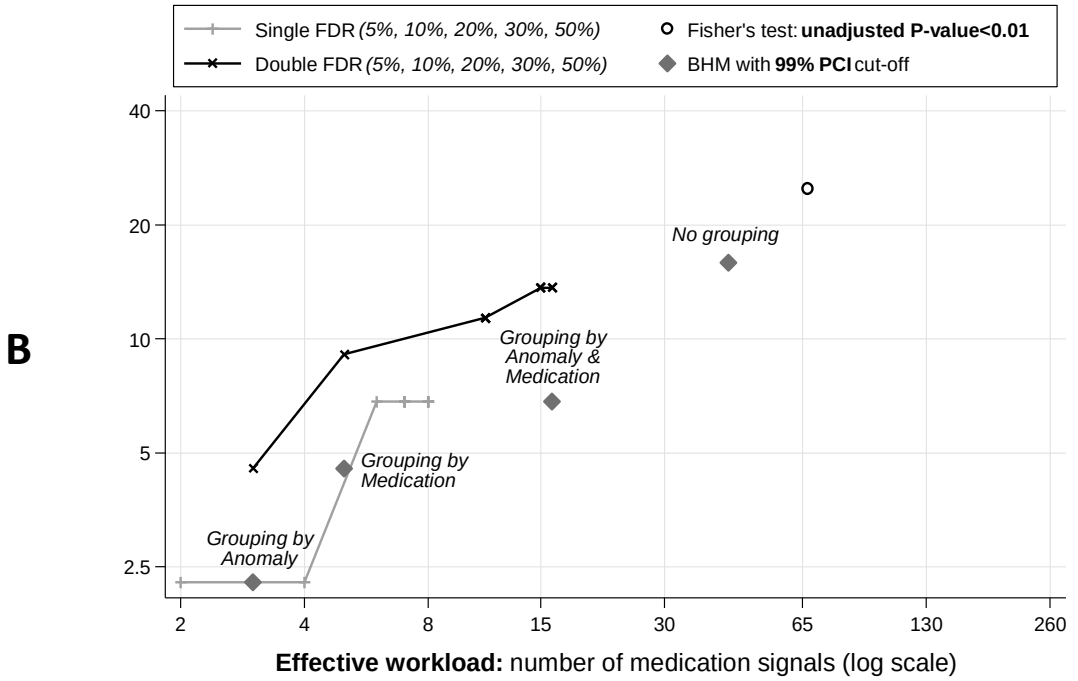
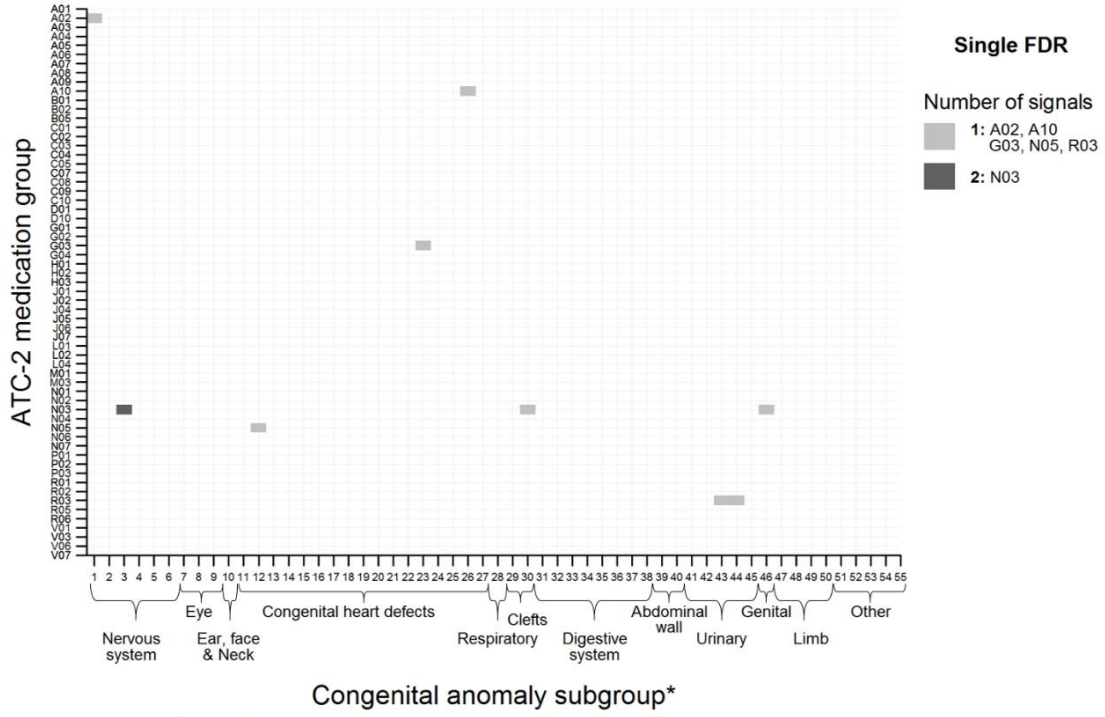
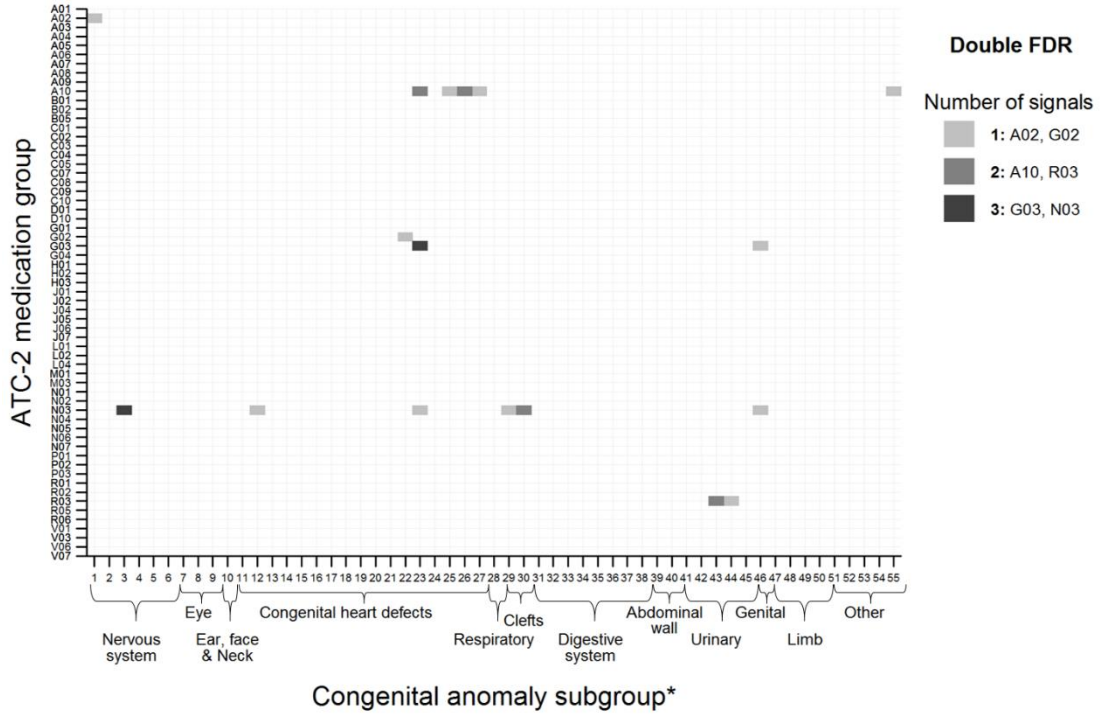


Figure 6.4. Identification rate (proportion of all "high risk" medications that are identified as signals) vs. effective workload: comparing the use of single and double FDR procedures with Poisson BHMs using (A) 95% PCI and (B) 99% PCI as a cut off for definition of signals for four types of grouping.



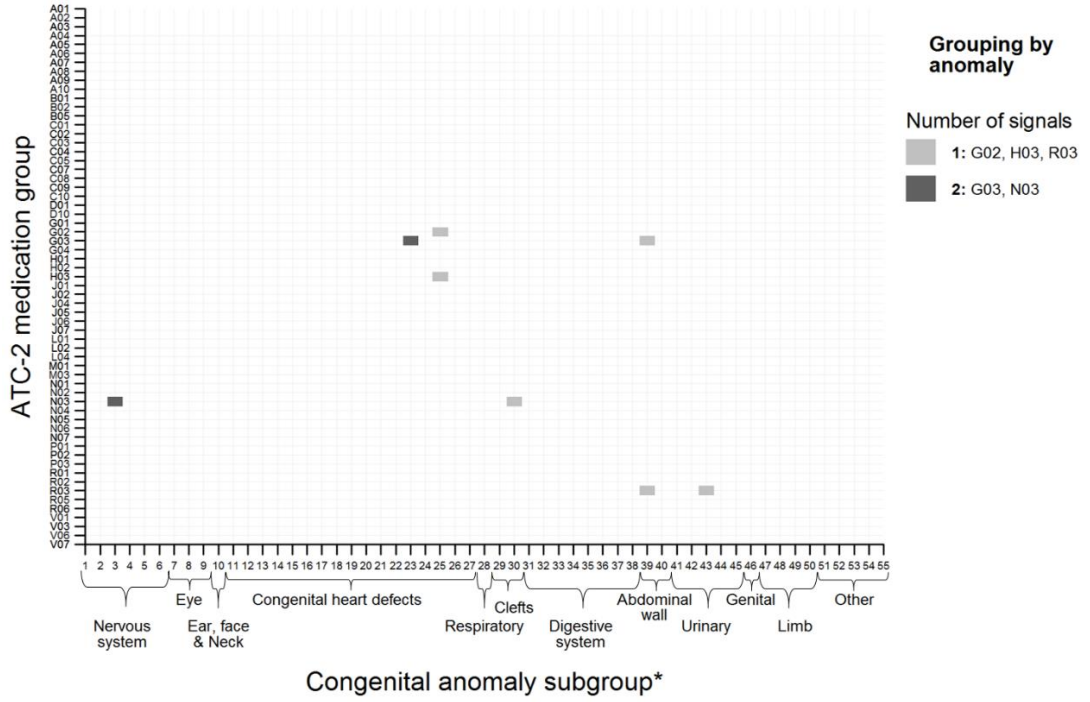
**See Table 4.4 for a full list of index numbers for all 55 CA subgroups*

Figure 6.5. Number of signals in each ATC-2 group of medications for each of 55 congenital anomalies, identified using a single FDR procedure with an FDR cut-off of 50%.



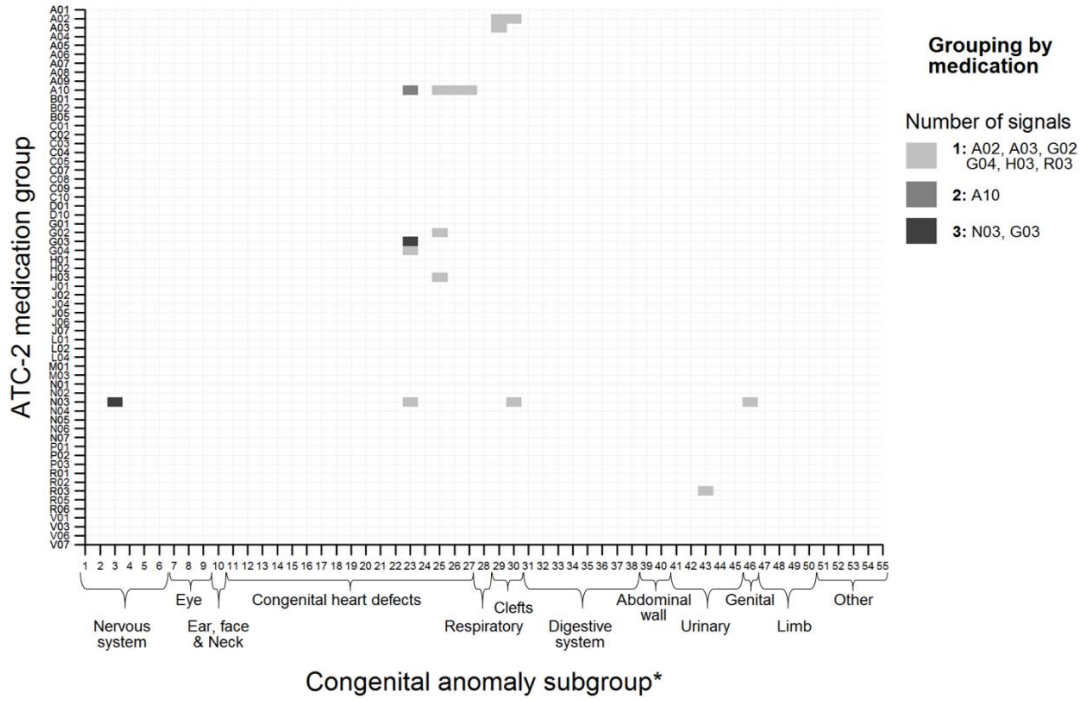
**See Table 4.4 for a full list of index numbers for all 55 CA subgroups*

Figure 6.6. Number of signals in each ATC-2 group of medications for each of 55 congenital anomalies, identified using a double FDR procedure with an FDR cut-off of 50%.



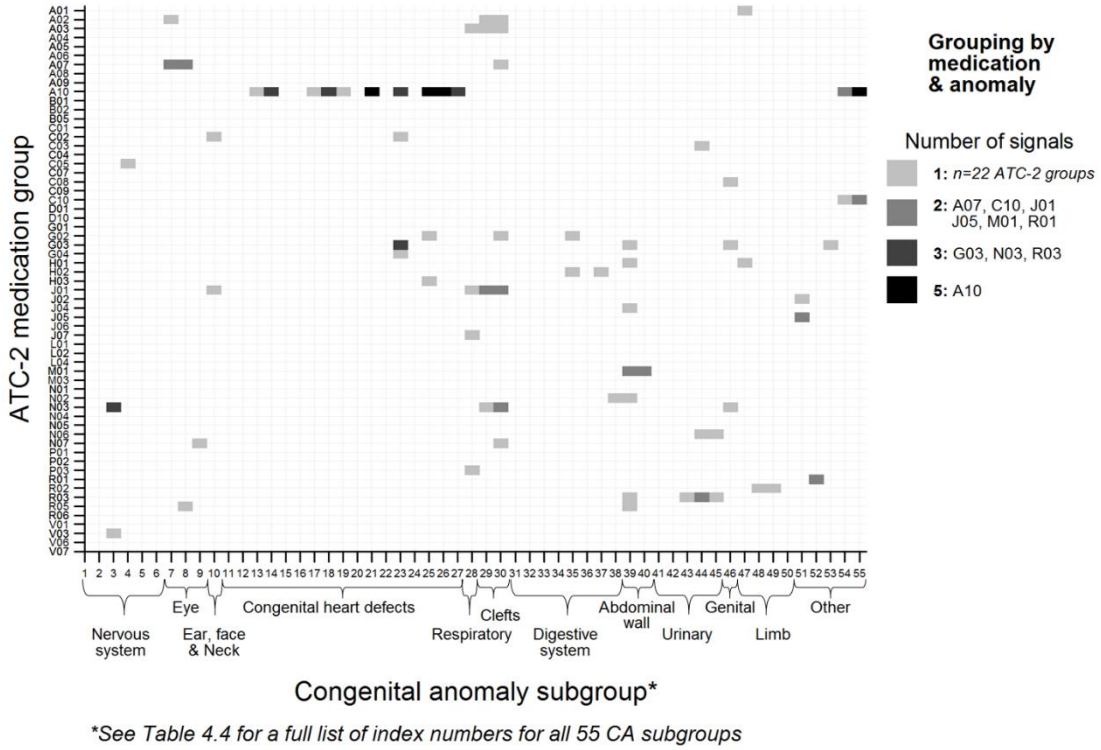
*See Table 4.4 for a full list of index numbers for all 55 CA subgroups

Figure 6.7. Number of signals in each ATC-2 group of medications for each of 55 congenital anomalies, identified using a BHM with grouping by congenital anomaly; 95% PCIs used to define signals.



*See Table 4.4 for a full list of index numbers for all 55 CA subgroups

Figure 6.8. Number of signals in each ATC-2 group of medications for each of 55 congenital anomalies, identified using a BHM with grouping by medications; 95% PCIs used to define signals.



*See Table 4.4 for a full list of index numbers for all 55 CA subgroups

Figure 6.9. Number of signals in each ATC-2 group of medications for each of 55 congenital anomalies, identified using a BHM with two-way grouping by congenital anomaly and medications; 95% PCIs used to define signals.

6.3.4. Different signals according to different approaches: the effect of shrinkage in Bayesian models

Differences between BHMs were apparent in the overall numbers of signals as well as which medication-CA combinations were identified as signals, as highlighted by the heatmaps in Figure 6.5 to Figure 6.9. This section presents examples of medication-CA combinations that were signals using some methods but not others, highlighting the effect of shrinkage in the BHMs. Estimates from combinations across the different Bayesian models and from a Fisher's exact test followed by a single or double FDR procedure are compared and discussed.

Example of shrinkage to the null: antiepileptic medications and atrial septal defect

Firstly consider an example of shrinkage to the null when using BHMs, where a signal attenuates towards the null due to the influence of other combinations in that group, meaning the group average is close to $log(PRR) = 0$ (i.e. no evidence that the medications in this group alter the risk of that particular CA). One example of this is the combination of the antiepileptic N03A medications with the CHD subgroup atrial septal defect. Figure 6.10 displays the estimated $log(PRR)$ and 95% PCIs for association of the 16 antiepileptic medications in combination with atrial septal defect according to 7 different methods of

analysis. Combinations considered a signal by any method are highlighted in black, whilst estimates in grey were not considered a signal according to that particular analysis. Note that a $log(PRR)$ was not estimated for combinations with a zero cell count when using Fisher's exact test (i.e. models with no P-value adjustment, single FDR or double FDR adjustment). However, the BHMs produced an estimate and PCI for all combinations, including those with a cell count of zero. Estimated $log(PRRs)$ for such combinations were sometimes non-zero due to the influence of other estimates in the group; however the 95% PCI for combinations with zero cell counts always crossed the null value of $log(PRR) = 0$. Therefore no combinations with zero cell count were identified as a signal in any Poisson BHM. The first two methods in Figure 6.10 are those without any adjustment for multiple testing: firstly signals identified using Fisher's exact test with no adjustment to the P-values, and secondly using individual Bayesian models with minimally informative priors. With no consideration of multiple testing, a frequentist analysis highlighted 3 N03A medications signals at the 5% significance level (i.e. using a P-value cut-off of 0.05). For individual BHMs, only 2 of these medications remained signals in combination with atrial septal defect; the estimate for N03AX11 was attenuated by the inclusion of a prior distribution for the $log(PRR)$ that was centred on a mean of zero. Note that the 95% PCI estimates for this model were generally a little wider than for other models. The third method in Figure 6.10 is the single FDR procedure, adjusting P-values across the whole dataset to account for multiple testing using a cut-off of 50%. This procedure did not consider any groupings but instead averaged estimates across all medication-CA combinations, and did not identify any N03A medications as signals in association with atrial septal defect. The double FDR method (again with a cut-off of 50%) considered the antiepileptics as a group, and in this case N03AG01 (valproic acid) was identified as a signal for atrial septal defect. This demonstrates how the double FDR method works in practice; the smallest P-value in the group here "passed" the first stage of the double FDR procedure, and all the N03A medications were therefore taken to the second stage. In this second stage of adjustment, a single FDR procedure was then applied across the set of medication-CA combinations from all the ATC3 groups whose minimum P-value passed the first stage of the double FDR. A BHM with grouping by medications gave the same result for N03A medications with atrial septal defect, with only N03GA01 being a signal. The last two methods in Figure 6.10 are BHMs considering groupings of CAs only and of CAs in combination with medication groups. For both these methods, Figure 6.10 demonstrates that the average estimates for combinations across the group of CHDs attenuated the signal for N03GA01 to the null.

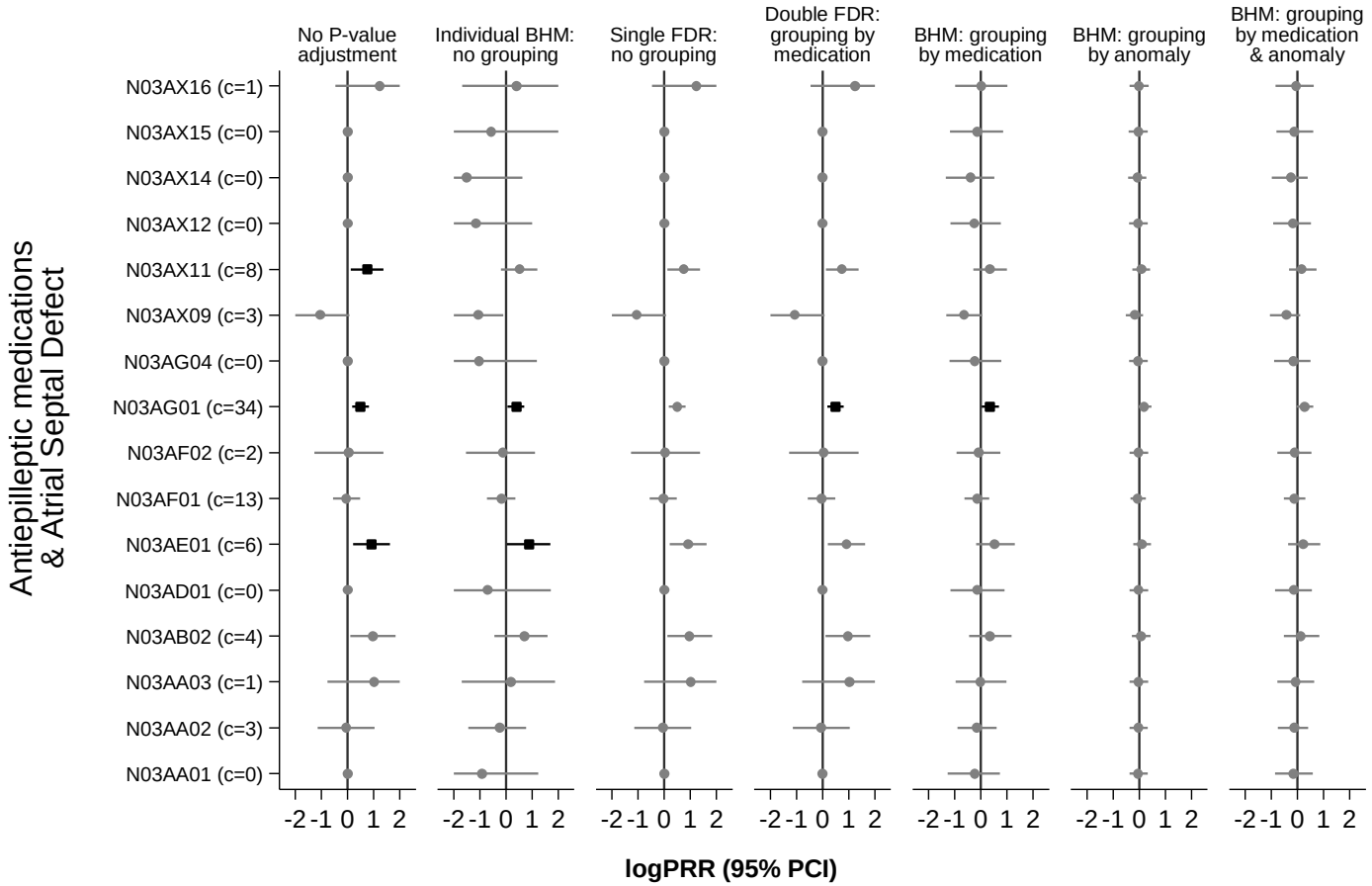


Figure 6.10. Estimated $log(PRR)$ and 95% PCIs for association of the N03A antiepileptic medications with atrial septal defect, according to seven methods of analysis. The number of exposures c for each medication in combination with atrial septal defect is shown in brackets after each ATC5 medication code. Estimates for all methods are truncated at -2 and 2 for visual purposes; those in black indicate combinations that are considered signals according to that method.

Example of shrinkage to the group mean (I): insulin medications and congenital heart defects

Shrinkage to the mean can also occur if one strong signal in a group has an effect on the estimates for other combinations in that group. Table 6.10 summarises the number of signals in the set of A10A insulin medications (19 ATC5 and 2 ATC4 coded medications) in combination with CHDs (17 subgroups) according to single and double FDR with a cut-off of 50%, and to the BHM grouping by both medications and CAs.

Table 6.10. Signals for group of A10A insulin medications (n=11) and congenital heart defect CAs (n=17) according to single and double FDR (50% cut-off) grouped by ATC3 codes, and BHMs (95% PCI cut-off) grouped by both ATC3 codes and CA groups.

Outcome	**Number of CAs**	**Congenital heart defect (CHD) subgroups**	**Number of signals per CA**		
			Single FDR: no grouping	Double FDR: ATC3 grouping	BHM: ATC3 & CA grouping
Signals in single FDR, double FDR *and* BHM	1	Patent ductus arteriosus (as only CHD in term infants)	1	2	4
Signals in double FDR *and* BHM	3	Atrial septal defect	-	2	3
		Ventricular septal defect	-	1	4
		Unspecified CHD	-	1	3
Signals in BHM only	6	Coarctation of aorta Pulmonary valve atresia Tetralogy of Fallot	-	-	1
		Common arterial truncus Single ventricle	-	-	3
		Transposition of great vessels	-	-	4
No signals for single FDR, double FDR or BHM	7	Aortic valve atresia/stenosis Atrioventricular septal defect Ebstein's CA Hypoplastic right heart Pulmonary valve stenosis Total anomalous pulmonary venous return Tricuspid atresia and stenosis	-	-	-
Total number of A10A-CHD combinations as signals			**1**	**6**	**27**
Total number of A10A medication signals in CHD subgroups			**1**	**4**	**11**

The medication A10AC01 (human insulin) was a signal for the CHD patent ductus arteriosus across all three of the methods in Table 6.10. Every combination including an A10A medication was taken to the second stage of the double FDR adjustment due to the minimum P-value for that group (the combination of A10AC01 with patent ductus arteriosus; $P=7.1x10^{-6}$) passing the first stage of FDR adjustment. All A10A medication-CA combinations were then considered in the second stage of P-value adjustment. Note that this procedure does not take groups of CAs into account, but averages across all CAs (i.e. not just the CHDs) when considering the statistical significance of each test. The double FDR method identified six signals of an A10A medication with a CHD, for four A10A medications across four CHD subgroups. When considering grouping by both medications and CAs using a BHM, there were 27 signals in the two dimensional A10A-CHDs set. This included at least one signal for every medication in the A10A group, in combination with 10 of the 17 CHD subgroups.

Another example where there was one combination that had an effect on the estimates for other combinations in that group was the A10A medications in combination with ventricular septal defect, where Figure 6.11 shows that the estimated $log(PRRs)$ shrank towards to the average estimate across this two-dimensional set. Note that Figure 6.11 shows there were four medications (A10AB, A10AB05, A10AC01 and A10AD05), where the estimated $log(PRR)$ had a 95% CI or 95% PCI that did not cross the null value of zero in any of these three models. Using single FDR adjustment, however, none of these combinations were considered signals. After double FDR, only one medication (A10AB05; insulin aspart) was a signal in combination with ventricular septal defect. In the BHM grouping by ATC3 medications and by CAs, all four combinations were flagged as signals. This highlights the way in which a BHM "borrows" information within each two-dimensional group; the model considers the evidence of an association to be stronger for these four combinations because they are similar CAs (i.e. CHDs) and belong to the same pharmacological subgroup of ATC coding (i.e. ATC3 group A10A). Figure 6.11 also illustrates how the large cell count for one or two CA-medication combinations can influence the group as a whole. The most common medications in combination with ventricular septal defect were A10AB05 (35 exposures) and A10AC01 (29 exposures) and for these combinations the effect of shrinkage appeared small, with the estimate and 95% PCI very similar to those obtained in the Frequentist analysis. Less common medications in this group, on the other hand, such as A10AD05 (n=4 exposures) exhibited more marked shrinkage towards the group mean and

smaller estimates for the $log(PRR)$. This demonstrates how combinations with fewer exposures (i.e. more limited information) have less influence on the average posterior distribution than combinations that are much more common.

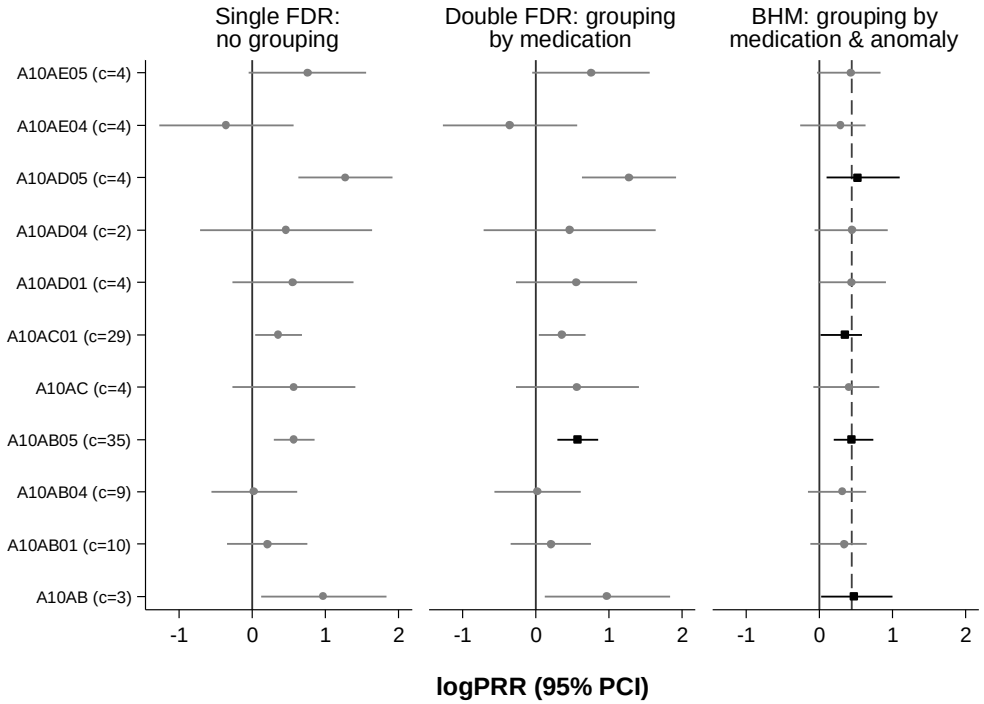


Figure 6.11. Estimated $log(PRR)$ and 95% PCIs for association of A10A medications with ventricular septal defect according to single FDR, double FDR grouped by ATC3 codes and a BHM grouped by ATC3 codes and CA groups. The dashed line in the BHM shows the mean $log(PRR)$ across the group of A10A medications and CHDs.

The shrinkage of estimates for A10A medications with patent ductus towards the average of all $log(PRRs)$ in this two-dimensional set is displayed in Figure 6.12. This is another example of one strong association in the group that was a signal across the three methods displayed. In this example, an additional signal was identified in the same group when using double FDR, and three extra signals when using a BHM grouping by both medications and CAs. Note that the BHM estimates for A10A medications in combination with this CA were again markedly affected by shrinkage, having narrower 95% PCIs than their corresponding frequentist 95% CIs. There were generally less exposures for A10A medications with patent ductus arteriosus than in the previous example of ventricular septal defect (which is the most common CHD), and in particular there were four medications with a cell count of only one exposure. For these medications the effect of shrinkage was clear, with the 95% PCIs being narrower and one of these medications (A10AE05) becoming a signal due to this shrinkage in the BHM with grouping by ATC3 medications and CAs. As mentioned

previously, combinations with a cell count of zero do have an estimated $log(PRR)$ in the BHM, but these always have a 95% PCI including zero (A10AD05, A10AC and A10AB in Figure 6.12). Another medication with a slightly higher count (A10AB05, n=5) was also a signal for patent ductus arteriosus in the BHM.

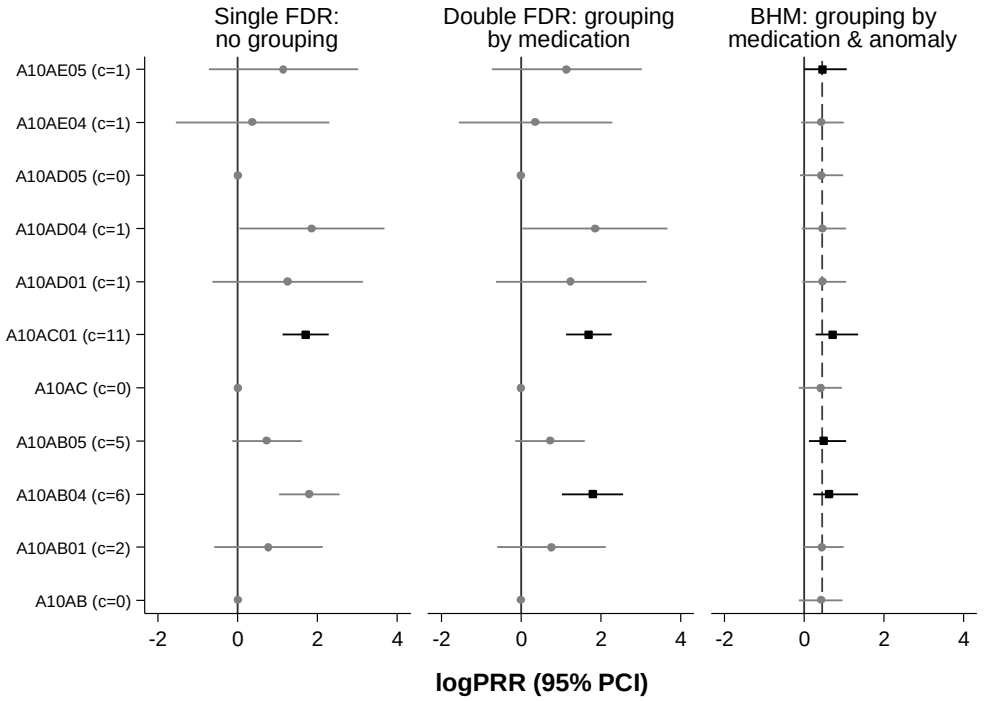


Figure 6.12. Estimated $log(PRR)$ and 95% PCIs for association of A10A medications with patent ductus arteriosus as only CHD in term infants, according to single FDR, double FDR grouped by ATC3 codes and a BHM grouped by ATC3 codes and CA groups. The dashed line in the BHM shows the mean $log(PRR)$ across the group of A10A medications and CHDs.

Example of shrinkage to the group mean (II): antiasthmatic medications and multicystic renal dysplasia

Notable shrinkage towards the group mean also occurred for groups where the data were very sparse; one such example is the combination of the urinary CA multicystic renal dysplasia with groups of medications for obstructive airway diseases. In Figure 6.13, estimates are shown for the four ATC3 groups within the ATC2 group R03 in association with multicystic renal dysplasia. There were no medication exposures for the medications in the two ATC3 groups R03B and R03C with this CA, so within these groups there were no reported signals according to any method assessed (although the BHM did produce estimates for these combinations). Another of the ATC3 groups R03D had limited information, with four of its seven combinations having no medication exposures. There was one medication within this group that had an estimate and 95% CI above zero in the

frequentist analyses, but this association did not remain a signal after either single or double FDR adjustment, and shrank toward the group average (which is the null in this case) in a BHM. In the group R03A there were more exposures but again no signals when using single or double FDR, despite there being three medications (R03AL01, R03AK06 and R03AC02) with an estimate and 95% CI above zero (Figure 6.13). When using a BHM with two-dimensional groupings, the two commonest combinations in the group R03A became signals as they shrank towards the group mean. The association for R03AL01, however, was not a signal in this method either, and the related $log(PRR)$ shrank towards the null with a 95% PCI that included zero.

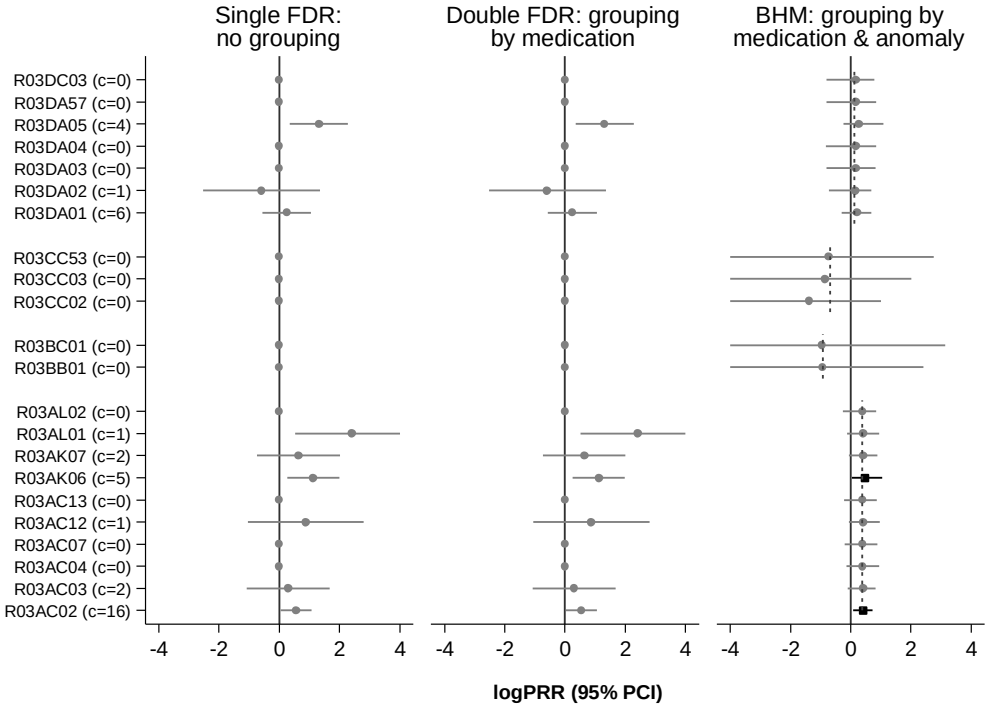


Figure 6.13. Estimated $log(PRR)$ and 95% PCIs for association of multicystic renal dysplasia with R03 medications according to single FDR, double FDR grouped by ATC3 codes and a BHM grouped by both ATC3 and CAs. Some lower 95% PCI limits are truncated at -4 for illustrative purposes. The dashed lines in the BHM show the mean $log(PRR)$ across each group of ATC3 codes for obstructive airway diseases medications in combination with the urinary CAs.

6.3.5. "Protective" associations in Bayesian Hierarchical Models

The number of significant "protective" associations according to each method is presented in Table 6.11, as well as the total number of combinations identified as signals (as shown previously in Table 6.9) for comparison purposes. Significant "protective" associations are those with a PRR<1 and a P-value passing the FDR procedure (for FDR methods), or a PRR and upper 95% PCI limit both below 1 (for Bayesian models).

Table 6.11. Number of signals and "protective" associations according to different methods of signal detection analysis investigated in chapters 5 and 6.

Type of model	Type of grouping	FDR or PCI cut-off	Combinations identified as signals		"Protective" associations		
			< 3 exposures [a]	**Total**	No exposures	1-3 exposures	**Total**
Fisher's test & Single FDR	No grouping	50%	0	**10**	0	0	**2**
Fisher's test & Double FDR	By medication	50%	1	**25**	0	1	**2**
Poisson BHM	No grouping	95%	59	**223**	29	19	**69**
		99%	2	**53**	4	9	**23**
Poisson BHM	By medication	95%	0	**21**	0	3	**14**
		99%	0	**7**	0	0	**3**
Poisson BHM	By CA	95%	0	**10**	0	1	**7**
		99%	0	**3**	0	0	**1**
Poisson BHM	By medication & CA	95%	44	**112**	19	16	**53**
		99%	1	**24**	1	5	**9**

[a] *No combinations with a zero cell count were identified as signals by any method*

Single and double FDR with a cut-off of 50% resulted in the smallest number of "protective" associations. When considering the total number of signals identified by each method, all BHMs resulted in a substantial number of combinations that were "protective" associations. As a proportion of the total number of signals, this was lowest for an ungrouped BHM using a 95% PCI cut-off; for this model there were 223 signals, but a further 69 (almost a third again, i.e. 31% of 223) of the remaining combinations were also statistically significant, i.e. had a PRR and upper 95% PCI limit below 1. As a proportion of the total number of signals, the highest number of "protective" associations was for BHMs grouping only by CAs (giving 10 signals and 7 "protective" associations) and BHMs grouping only by medications (21 signals and 14 "protective" associations). Note that both "protective" associations for single FDR in Table 6.11 were for the CHD atrial septal defect in combination with G02CA and with N02BE05, and that atrial septal defect also had a strong signal for the sex hormone G03DB01.

6.4. Discussion

In this chapter, BHMs were applied to EUROmediCAT data in order to assess the effect of directly modelling group effects for medications and/or CAs in signal detection analyses. The way in which the groups were specified was the main source of prior information in these Bayesian models. Use of BHMs were compared to the multiple testing adjustment procedures investigated in the previous chapter (using the same dataset). The comparison was made to results from a single FDR, as this is the method currently used by EUROmediCAT for signal detection purposes, as well as the newer double FDR method, which showed some improvement over the single FDR in the previous chapter.

6.4.1. Use of Bayesian hierarchical models to detect signals of teratogenic medications in EUROmediCAT data

Overdispersion and the independence assumption in count models

An important assumption of the Poisson and a negative binomial distributions is that events in the data occur independently, i.e. that the exposures are independent from each other. Independence of reports is an underlying assumption in the calculation of confidence intervals for any signal detection method reporting a PRR (or ROR), and in practice for SR data it is often the case that a single individual case safety report may involve multiple medications and multiple AEs. Violation of this independence assumption can bias the estimation of the variance for the PRR, in turn affecting the number of combinations that are identified as signals. If analyses are done using counts of individual case safety reports, rather than numbers of medication or AE reports, violations of this assumption may be minimised [European Medicines Agency, 2012]. In EUROmediCAT data, a malformed foetus often has more than one CA and/or has been exposed to more than one medication, and can therefore contribute exposure counts to multiple medication-CA combinations in the dataset. In the data for this thesis, there was an average of 1.2 CAs and 1.5 medications per pregnancy (when looking at those subgroups and medications being monitored for signal detection analyses, see section 4.4.2). This means exposures are unlikely to be fully independent, since a count in one cell may refer to the same malformed foetus as an exposure to a different medication or CA counted in another cell. This may have led to overdispersion in EUROmediCAT data; in addition to a Poisson model, the negative binomial distribution was therefore also used to assess if overdispersion was present, and whether this might be accounted for by use of this more flexible model for the exposure counts.

In addition to cases having multiple CAs and multiple medications, there were other correlations in this data structure. Certain CAs and medications, for example, may be more likely to co-occur within pregnancies, even across the groupings used. Similarly, exposure to a certain medication may increase the likelihood of exposure to another medication, for example it is common to take several different asthma medications together. On the other hand, it may be the case that if a woman is taking one particular medication, then it is very unlikely that she will have been exposed to any other medications in that group, i.e. medications within a group may be mutually exclusive. These kind of situations could introduce specific dependence structures within the data that are not necessarily easy to quantify across the whole dataset. These would need to be considered on a case-by-case basis, which isn't realistic in practice for large databases like EUROmediCAT.

However, despite this potential for overdispersion and other dependencies in the data, there was little difference in the results when modelling the data using a Poisson or a negative binomial model (i.e. when assessing the effective workload, "high risk" proportion and identification rate). Furthermore, the estimated dispersion parameter in negative binomial models tended towards a large value for various choices of parameters for the prior distribution, indicating that inclusion of such a parameter did not materially affect or improve the model fit. The Poisson model, which is more parsimonious and easier to implement, was therefore used throughout analyses in the rest of the chapter.

Information sharing by groupings of either medications or CAs

Two of the BHMs considered in this chapter used discrete groupings for only one variable, i.e. for the medications or the CAs but not both. In these models, the variable that was not grouped was given an overall group prior distribution, such that there was also some level of information sharing within this variable. For example, when grouping only the medications (using their ATC3 codes) the CAs were given a common prior distribution within each ATC3 group, implying a common distribution of effects for each ATC3 group of medications across all of the CAs. However, if a similar distribution of effects for a group of medications would not be expected across all the CAs, then it might have made sense to allow for a different distributions within each CA. For example, certain antiepileptic medications are known to be teratogenic for certain CAs, but would not be expected to have similar effects across all of the different types of CAs; in general, a single teratogen may increase the risk for a number of different CAs, but would not be associated with an increased risk of all CAs. However, this opposite approach would have introduced a large number of additional parameters, therefore likely leading to over-parameterisation in the

model (e.g. additional parameters for the mean and variance for a further 54 prior distributions would be required to give each of the 55 CAs their own distribution of effects for each groups of medications if grouping by ATC3 codes).

Protective associations in Bayesian hierarchical models

The purpose of signal detection analyses is to screen for potentially harmful teratogenic medicines; any "protective" associations identified are not, therefore, flagged as signals or recommended for further examination. In addition, the power to detect protective associations is low in these analyses due to there being no healthy controls in the data. However, some "protective" associations did arise in the analyses, and there are a number of potential reasons for this. Firstly, these may indicate a true potentially "protective" effect of a medication; for example, if an insulin medication is demonstrating a "protective" effect this may indicate that it is providing better glycaemic control (i.e. treatment of the underlying disease) in comparison to other medications, which may in turn affect the risk of that foetus having a particular CA. Note that this is only in comparison to other medications and CAs, rather than a potentially preventive effect in general. On the other hand, it is unlikely that the "protective" associations encountered in these analyses present true treatment effects, as there are other potential factors that must be considered. In any signal detection analysis using disproportionality measures, frequently reported outcomes can make it difficult to detect signals relating to the medication in question. In the context of SR databases, particularly extreme signals can lead to the PRR being less than one for combinations of that medication with other AEs, as well as potentially affecting the detection of signals of that AE with other medications [Waller et al., 2004]. For example, an extreme signal may be for a particular combination of a harmful medication A with a common outcome (outcome 1), then in examining the association between that medication and a different outcome (outcome 2). Therefore, when considering everyone who has taken medication A, a large proportion of the these will have outcome 1. When a second outcome is then considered, those on medication A and with outcome 1 are now the "controls" in the association of medication A and outcome 2, meaning that outcome 2 will appear less frequent amongst those taking medication A compared to those who are not taking medication A. This could make it appear as though there is a protective effect of the medication on outcome 2.

For EUROmediCAT data, if there is a particular medication-CA combination where the medication has a high prevalence, then the *expected* cell counts for combinations of other CAs with that medication (or, likewise, other medications with that CA for a common CA)

will be driven up by the larger marginal total, which is being dominated by the common medication. If there is a signal (with a PRR>1) for this particular medication-CA combination, this may in turn cause one or more "protective associations" of that medication with other CAs. This is because the expected cell counts are more likely to be higher than the observed cell counts for the other combinations also including that particular medication. This issue is highlighted by the fact that all but one of such "protective" associations that arise in models for EUROmediCAT data were for CAs or medications that were signals in other combinations. For example, the insulin medications (ATC3 group A10A) are known to be associated with an increased risk of CHDs, and many of these combinations did show up here as signals (across all methods considered). This might be an example of confounding by indication, whereby the medication has been prescribed to treat a disease (i.e. diabetes) that is associated with the outcome (CHDs). Maternal insulin-dependent diabetes has long been known to be an important risk factor for congenital malformations. Mills [2010] presented a review highlighting consistent evidence of increased risks for a range of CAs amongst infants of diabetic mothers, including cardiovascular, musculoskeletal, genitourinary and other types of malformations. Signals of insulin medications are therefore expected to occur in any unadjusted analysis that does not adjust for information on maternal illnesses, such as the signal detection analyses considered here. It has been suggested that the use of insulin analogues in pregnancy may be associated with a higher risk of CAs in comparison to human insulin; however, a recent literature review found that there was no evidence of such an increased risk for insulin analogues [de Jong et al., 2016b]. A EUROCAT study found that CHDs occurred around twice as frequently in diabetic pregnancies compared to other CAs, and these accounted for the majority of the excess risk of CAs amongst exposed pregnancies [Garne et al., 2012a]. Garne et al. also found that oro-facial clefts and limb defects were less likely to be associated with diabetic pregnancies, relative to other CAs. Likewise, in this thesis, a number of the insulin medications were also "protective" associations with oro-facial clefts and limb CAs. For these groups of CAs the expected counts used in BHMs were driven up by the higher prevalence of these medications overall (due to high numbers amongst the heart defect subgroups), therefore making the observed values for other CAs much lower than the expected (Appendix Table C1). An example of this situation is for the medication A10AC01 (human insulin), which was seemingly "protective" for oro-facial clefts and limb defects, relative to other CAs. This is likely to be driven by the high numbers of exposures to the combination of this medication in foetuses with CHDs (see Appendix Figure C1). It is

important to note that the presence of such “protective” associations does not mean that either insulin or maternal diabetes are likely to be protective for oro-facial clefts or limb defects, but rather reflects that the comparison here is of the effect of a particular medication compared to *other medications in the database*. It is not possible to quantify the effect of a medication compared to any other possible medication or, indeed, compared to the effect of taking no medications at all on the risk of any particular CA. Likewise, the risk for any specific CA in the data is made in comparison to all *other CAs in the database*, and is not a comparison to a pregnancy without a CA. It should therefore be again be noted that a “protective” association is not protective per se, but may be indicative that a particular medication is less of a risk than another medication in the database (that may be teratogenic). As such, the insulin medications may be thought of as carrying less of a risk for limb and cleft CAs when compared to their risk for CHDs.

Nearly all the protective associations in these analyses included a medication or CA present in another combination that was identified as a signal (Table 6.11) ; the only exception to this was the digestive system CA ano-rectal atresia and stenosis in combination with the medication A03FA01 (metoclopramide; a propulsive medication for functional gastrointestinal disorders). This combination had a cell count of only 2, and was present as a signal only in a BHM with grouping by both medications and CAs. Appendix Figure C2 displays the estimates for combinations of the 8 digestive CAs and 3 medications within this medication-CA set according to the single and double FDR models as well as a two-dimensional BHM. The “protective” association for A03FA01 in the BHM with grouping by CA and medications was influenced in this example by the group mean, which was a $log(PRR) < 0$ due to cell counts of zero for the majority of combinations (i.e. for which there were no exposures to that particular medication and CA combination) within this group. Since, by definition, the marginal total for each medication and for each CA must both be at least 3 to be included in the signal detection analysis dataset, the expected cell counts will always be greater than zero. Therefore comparing observed and expected values for combinations with a cell count of zero will always result in a PRR<1 (and a $log(PRR) < 0$, as in Appendix Figure C2). Whilst the medication and the CA in this protective association were not involved in any other signal combinations, there was a signal in common with another medication in this group for a different type of CA (A03FA with the respiratory system CA choanal atresia), and there were four signals for digestive system CAs with medications in other groups (G02CA01 and H02AB01 with diaphragmatic hernia; H02AB01 with Hirschsprung's disease; N02BA01 with oesophageal atresia).

6.4.2. Comparison of Bayesian hierarchical models and false discovery rate procedures for signal detection in medication safety data for congenital anomalies

The double FDR identified more "high risk" signals (increased "high risk" proportion and identification rate) for comparable effective workloads as BHMs. Ungrouped BHMs and those grouped by both medications and CAs sometimes identified greater absolute numbers of "high risk" medications, but this was at the expense of a substantial increase in effective workload. When also taking into account the increased computational time and effort involved in the implementation of BHMs compared to the double FDR method, it is therefore recommended that the double FDR method be used in practice for the detection of signals of teratogenic medications using EUROmediCAT data.

Different signals according to FDR and BHM approaches

A selection of results were presented in greater detail in section 6.3.4 to highlight the way in which the different approaches to signal detection analyses and the choice of grouping within BHMs resulted in different sets of combinations being signals. In BHMs, shrinkage was expected to have a similar effect to adjustment for multiple testing in a frequentist setting in terms of a reduction in the overall false positive rate. Information considered within and around each cell of interest led to adjustment of PCIs and therefore the resulting set of signals. When other combinations in a group contain additional information to that of the combination of interest, this can reduce the significance level for results that are likely to be false positives, whilst enhancing the significance of those likely to be true associations. For combinations with smaller cell counts and/or lower marginal totals, more of a shrinkage effect was expected, since the posterior for such combinations is influenced more by the prior and less by the likelihood function (i.e. the actual data) due to there being less information in the data itself. How useful this is depends, of course, on the strength of information in the other cells in any set (i.e. whether there is actually useful information within that particular group).

One example that was presented of differing signals from the different approaches was the double FDR and BHM with grouping by ATC3, which both identified the combination of valproic acid with the CHD atrial septal defect as a signal. However, this combination was not a signal using single FDR or BHMs with CA grouping (Figure 6.10), implying that there was a null effect on average across the antiepileptic medications in combination with the CHDs as a group. A significantly increased risk of atrial septal defect has previously been demonstrated for valproic acid monotherapy, whether compared to a control group taking

no antiepileptic medications or a control group taking any other antiepileptic medication monotherapy [Jentink et al., 2010]. This highlights how different groupings can have different conclusions, even for a known teratogenic medication such as valproic acid.

In methods that did not consider grouping CAs or medications in the analysis stage, combinations with a cell count of only one or two were not considered signals even if they showed a significant association, since estimates based on such small counts are not judged sufficient information on which to base a conclusion and can be more likely to be false positive associations. One of the potential advantages of using BHMs here was that it might strengthen the analysis of combinations with low cell counts by using information in the surrounding cells, allowing them to potentially be included in the set of resulting signals. However, no combinations with a cell count lower than three were identified as signals in BHM with discrete grouping only by medications or CAs. In the BHM with discrete groupings by both CAs and medications, 34 medications over 44 combinations with a count of one (n=26) or two (n=18) were identified as signals (Table 6.9), including 3 "high risk" medications, 18 "low risk" medications and 13 medications with no known risk category. If these combinations had been excluded from the list of potential signals (as in FDR methods) there would be an improvement in the proportion of signals that were "high risk" ("high risk" proportion increases from 14% to 17%) but a decrease in the percentage of all potential "high risk" medications being identified (identification rate decreases from 23% to 16%). This is because three out of the 10 (see Table 6.9) "high risk" medications in the set of signals were combinations with a frequency of only one or two in the data.

The double FDR 50% and a BHM using the same grouping by ATC3 medications (in both cases averaging over all CAs) resulted in 16 and 15 medications being identified as signals, with a total of 25 and 21 combinations, respectively (Table 6.9). However, two more "high risk" medications (6 vs. 4) were identified by double FDR. Ten medications, including 15 combinations, were signals according to both methods. Of the remaining signals for these two methods:

- **5 medications identified by a BHM with discrete grouping by medications were *not* signals when using double FDR.** One was a thyroid hormone H03AA01 in combination with the CHD ventricular septal defect. There has been some evidence of an increased risk of birth defects (including CHDs) following the use of thyroid medications in pregnancy in a Danish population based cohort study [Andersen et al., 2013]; however, this particular thyroid medication H03AA01 is a "low risk" category A medication in the Australian risk categorisation system database. Another was the antacid A02AD01 in

combination with Oro-facial clefts (cleft palate and cleft lip ± palate subgroups). Evidence regarding this medication is limited, although antacid use has been associated with a *decreased* risk of clefts in a case-control study of the treatment of nausea and vomiting during pregnancy and the risk of various common non-cardiac CAs in the US [Anderka et al., 2012]. The remaining three medications were A03AD02 with cleft lip ± palate, G02CA with ventricular septal defect (although this medication was a signal in combination with tricuspid atresia & stenosis, another CHD subgroup, when using double FDR) and G04BX with atrial septal defect. These medications all had unknown risk category and a lower 95 % PCI for the PRR close to 1, i.e. they were close to the threshold of being defined as a signal.

- **On the other hand, 6 medications (9 combinations) were signals using double FDR but *not* in a BHM grouped by medications.** These were the 9 least frequent combinations in the set of signals resulting from the double FDR with ATC3 groupings (each with an exposure count of 8 or less). Perhaps most notable of these is the combination of Anencephaly with misoprostol (A02BB01), a "high risk" category X medication for which there has been evidence of an increased risk associated with birth defects such as brainstem injuries [Vauzelle et al., 2013]. This medication was not a signal in any of the BHMs that considered grouping. The ATC3 group A02B included ten distinct ATC5 coded medications of which eight were "low risk" (category B) medications, one did not have a risk category assigned, and the final "high risk" (category X) medication was A02BB01. The majority of combinations of any A02B medication with one of the 55 CAs in the analysis had a zero count. For the BHM with groupings by ATC3, the average PRR across the A02B group of medications was 0.9, i.e. a reduced risk across the CAs when compared to other medications in the data. The majority of medications in this group were "low risk" and, in addition, this combination had an exposure count of only three, hence it is not surprising that using a BHM shrinks the estimate for the one "high risk" medication to the null. Of the other six medications that were signals in double FDR but not a BHM with ATC3 grouping, there were two antiepileptic medications in the "high risk" category (N03AA02 and N03AF01). The other three such signals were "low risk" medications (A10AB04, N03AX14 and R03CA02) and one had unknown risk category (G02CA).

Masking in signal detection analyses

The PRR is a ratio and therefore depends on both the numerator, which includes only reports involving the medication of interest, and the denominator, which includes all

reports involving other medications. The numerator may suffer from biases due to issues with reporting for the medication of interest. For example, if there is excessive reporting for a common medication (e.g. due to media influences) then the overall rate for this medication will be inflated, and other associations may be concealed. If the CA of interest is significantly associated with medications other than that under consideration, the denominator will be inflated as the background reporting rate (i.e. the marginal total) for that medication and event in combination is increased. In practice this means that the expected count will be higher for other medications in combination with the CA, so the observed to expected ratio may then be close to one, even if there is a true signal for another medication. This statistical issue is known as masking and has previously been discussed in a number of articles on data mining techniques for pharmacovigilance [Gould A. L., 2003, Almenoff June et al., 2005, Hauben et al., 2005]. Confounding by co-reported medications may also occur if two medications are frequently prescribed together but only one causes the CA of interest [Hauben et al., 2005]. For EUROmediCAT data, masking may also occur due to the use of malformed controls if a proportion of the control group is related to the medication of interest. The most common type of CA in the EUROmediCAT database, for example, are CHDs, which accounted for 35% of all malformed foetuses (Table 4.5). Of all recorded CAs in the data, 17% had ventricular septal defect, the most common CHD. If a particular medication is reported frequently in combination with ventricular septal defect, this could mask the association of another (less common) CA with the same medication, as the expected count for such a combination will be inflated by the high prevalence of this medication with ventricular septal defect.

Studies have demonstrated that the removal of a masking effect may help lead to new signals of public health relevance being discovered [Gould A. L., 2003, Pariente et al., 2012]. It is also thought, however, that significant masking is not common in large SR databases, and where present it mostly affects rarely reported AEs [Zeinoun et al., 2009, Wang et al., 2010, Maignen et al., 2014]. Various approaches to deal with masking in SR databases have been described and assessed by Wang et al. [2010], who concluded that considerable resources could be required to unmask only a small number of masked signals, that may have only weak evidence for additional causal associations. Another potential issue with identifying and dealing with masking is that the type I error (false positive) rate can be artificially inflated [Maignen et al., 2014]. Attempts to unmask associations should therefore be based on prior information on known specific associations that may be thought to have an important masking effect, rather than using routine and

computationally intensive methods to try and account for masking in the analysis. After signal detection it may be beneficial to perform a secondary analysis to assess the effect of unmasking in such cases where there is a strong previously known association. This raises the question of how to deal with previously known associations in updated signal detection analyses (see section 7.4.2 for further discussion on this point).

6.4.3. Strengths and limitations of EUROmediCAT data

A major strength of EUROmediCAT data is the detailed and standardised coding of the CA outcomes across the registries due to the existing EUROCAT network upon which EUROmediCAT is based [Boyd et al., 2011], as well as detailed information regarding medications taken during the first trimester of pregnancy. Good agreement between the medication that was actually used and that recorded in one EUROmediCAT registry has also been demonstrated in a study that used additional data sources compared to those used to contribute data to EUROmediCAT [de Jonge et al., 2015a]. One weakness of EUROmediCAT data is that there is no information available regarding the dosage of medications. In addition, the timing of exposures could not be confirmed in a number of cases in the data for this thesis. In particular, a high proportion of recorded cases for the Polish registries could not be confirmed as first trimester medications, and were therefore excluded from these analyses. This will have resulted in a loss of power and may be a source of potential bias. However, similar distributions of types of CA were present in the Polish registries for those with confirmed first trimester medications and those with unknown timing, such that those cases that were included were considered not dissimilar to those excluded due to unconfirmed medication timing (and any related potential biases minimised). Data cleaning for timing of exposure meant that all included cases were supposedly confirmed first trimester exposures, however it is not possible to know that the mothers definitely took the medication during the critical period for development of each specific CA [Czeizel, 2008]. Another potential limitation is that there is known under ascertainment of some medications in EUROmediCAT data [Bakker and Jonge, 2014, de Jonge et al., 2015b] and this may reduce the sensitivity of any signal detection analyses. In addition, ATC coding changes over time were taken into account and updated as far as possible; however, it is feasible that some changes were missed and this this may potentially have led to signals being undetected if the exposed cases were divided across multiple ATC codes as a result.

6.4.4. Summary and conclusions

In this chapter, the use of Bayesian models showed some interesting properties in terms of shrinkage to group means and identification of different medication-CA combinations as being signals. However, in general the BHMs did not produce enough of an improvement in terms of the signals identified to justify the increased complexity and workload of implementing these models. Only the BHM that grouped combinations discretely by both medications and CAs identified more "high risk" medication signals compared to the double FDR, but this came at the cost of a much larger resulting workload of signals to follow up in the next stage of the signal management process. Overall, no BHM performed better than the double FDR[1], which is relatively straightforward to implement and is therefore recommended for use in future signal detection analyses for EUROmediCAT data.

[1] See Table 5.4 for details of the 16 medication signals identified by the double FDR a 50% cut-off and ATC3 groupings.

Chapter 7: Synthesis of thesis findings

7.1. Introduction

This thesis investigated whether taking existing hierarchical structures in the coding of CAs and medications into account would lead to more accurate statistical models in the analysis of CA prevalence data. Data sources, statistical methods, results and conclusions for these different approaches to the improvement of CA surveillance methods were presented. These were applied to two datasets; one including counts for CA prevalence from population-based registries across Europe, and the other coming from those registries that had additional information regarding first trimester medication use. In this chapter, the findings from this thesis are summarised and discussed, including methodological questions and potential areas for future research.

7.2. Summary of work presented

Chapter 1 introduced the thesis, comprising a background to the main themes of CA surveillance, Bayesian analysis and the use of hierarchical models. **Chapter 2** presented the overall aim and specific objectives of this thesis. **Chapter 3** investigated the use of BHMs for the routine analysis of CA prevalence data. Specific types of CA were explored that illustrated situations in which grouping CAs together was considered potentially useful. BHMs demonstrated some interesting properties with regards to shrinkage to group means when sharing information within EUROCAT defined groups of CAs. However, overall there was little difference in results when compared to models that analysed each CA individually. In addition, BHMs were more complex and time consuming to implement, and some models showed a lack of convergence. The next three chapters moved on to the analysis of medication use during first trimester of pregnancy and the associated risk of CAs; **chapter 4** reviewed methods for signal detection in SR databases and in CA data, using these to identify two approaches that might improve the detection of teratogenic medications in EUROmediCAT data. The EUROmediCAT dataset for use in these analyses was then described, and the Australian risk categorisation system for prescribing medications during pregnancy was presented as a way of independently identifying the number of "high risk" medications that each method was picking up. **Chapter 5** considered different FDR procedures that incorporated information about groups of similar medications and/or CAs when determining the statistical significance of the test for each medication-CA combination, whilst **chapter 6** assessed the use of BHMs to directly model

the potential group effects. Throughout this thesis, various applications of BHMs did not produce enough of an improvement to justify implementing such models. The double FDR method grouping medications by ATC3 level codes performed better than BHMs and other FDR methods (including the currently used single FDR) in terms of metrics based on the number of "high risk" signals identified by the Australian risk categorisation system. In conclusion, it was recommended that (i) when using EUROCAT subgroups for analysis, considering each CA separately remains an appropriate method for the detection of potential changes in prevalence by relevant surveillance systems, and (ii) the double FDR procedure grouping medications by ATC3 level codes should be used in future signal detection analysis for CA data using ATC coding for medications and EUROCAT coded CAs.

7.3. Methodological considerations

7.3.1. Use of Bayesian hierarchical models

BHMs were applied to both CA prevalence and first trimester medication data, however in both scenarios the models investigated in this thesis did not produce enough of an improvement over other methods to justify their use. When assessing recent ten-year trends, it is likely that there were not sufficient data points to support the additional random effect parameters required to group CAs together, especially for models that also included random effects for registry. In addition, for small groups of related CAs (e.g. the three NTD subgroups) there were not enough levels for random effects parameters to be reliably estimated. Indeed, estimation of random effects parameters with less than 5 levels have been shown to perform poorly in BHMs, resulting in a lack of convergence and over-parameterisation [Greenland, 2000, Gelman Andrew and Hill, 2007]. On the other hand, there was too much heterogeneity in the prevalence of EUROCAT subgroups for more than 5 CAs (such as the CHDs) to be grouped together as random effects in a BHM, even for those CAs within the same organ/body system.

Application of BHMs to both CA prevalence and medication data also demonstrated the effect of shrinkage. In signal detection analyses using EUROmediCAT data, estimates were affected by the groupings used such that medication-CA combinations were signals only in certain models. In some cases this shrinkage appeared potentially useful, for example there were signals for 4 insulin medications in combination with the CHD ventricular septal defect when using a BHM with groupings by both medications and CAs, however single FDR resulted in no signals and double FDR only one signal in this group (A10A medications in combination with ventricular septal defect; see Figure 6.11). Although signals for insulin

medications are thought to be a marker of the increased risk of CHDs associated with maternal diabetes (rather than an indication that these medications themselves are teratogenic), a signal detection method should identify the type of associations where further investigation is worthwhile. On the other hand, shrinkage in BHMs also attenuated the estimated risk for some known teratogens; for example the known association between valproic acid (N03AG01) and the CHD atrial septal defect was a signal after double FDR, but in a BHM with groupings by both medications and CA this combination was not a signal due to an overall average null effect across the group of antiepileptic medications and CHDs (Figure 6.10). It is not ideal for a method to miss a known teratogenic signal when another method picks this signal up using the same data. This example highlights the potential for shrinkage in Bayesian models to have a "self-fulfilling prophecy", where if there are truly different trends across "similar groups of CA" then shrinkage will make them "disappear". This may be problematic since CAs that are grouped together are not necessarily similar in terms of their aetiology, risk factors or temporal trends.

All BHMs in this thesis used minimally informative prior distributions for parameters, whereby the means followed a Normal distribution centred on a $log(PRR)$ of zero, and standard deviations followed a uniform distribution. The sensitivity of these priors was tested for all models to ensure that they were robust to changes in parameter values, such as the choice of limits for the uniform distribution or the variance of the Normal distribution. In the analysis of EUROCAT prevalence data, BHMs often demonstrated overfitting and a lack of convergence. BHMs for EUROmediCAT data generally showed good mixing and convergence, although high levels of autocorrelation were present and long chains with a large thin were therefore required. The thinning of chains in MCMC samples is frequently used in practice, although greater precision in estimates is obtained by working with chains that have not been thinned [Link and Eaton, 2012]. However, from a practical perspective, computing memory and storage limitations must also be taken into consideration; for very large chains that are not thinned, post sampling calculations (i.e. model assessment and summarising posterior distributions) can impose a substantial computational burden.

7.3.2. Treatment of registry in analyses

In the analysis of CA prevalence, models are frequently adjusted for the effect of registry (see section 3.2) since prevalence often varies in the different registries. This can reflect true differences in prevalence and trends within different regions and countries, but may

also be due to differences in coding practices and/or information sources in each region. Adjustment for registry also accounts for the fact that registries have differing maternal age structures and can contribute to different time periods in EUROCAT data. Since CA prevalence data are expected to be prone to regional differences, it was important that the effect of adjusting for registry was assessed in models in chapter 3. For medication data, however, a teratogen is expected to act in a similar way regardless of where it is taken. One aspect that may vary between registries within EUROmediCAT data is that of varying usages and/or availability in different countries for certain medications. As there are very small numbers for many medication-CA combinations, the best approach to an ongoing signal detection process is considered to be investigation of any potential registry effects at a later stage in the analysis. Indeed, this is part of the next step of the signal detection process, where the first step in the follow-up of any potential new associations resulting from the signal detection analyses is to perform data validation, which includes the adjustment of estimates for confounding by registry [Given et al., 2016].

7.3.3. Importance of grouping choices for medications and congenital anomalies

The main objective of this thesis was to investigate the use of groupings of ATC codes and/or CAs in the analysis of EUROCAT and EUROmediCAT data. For FDR methods, the different types of ATC groupings had less of an effect on the results than the type of FDR procedure that was used. For grouping in BHMs, minimally informative parameters were used in prior distributions and the main source of "informative" prior information therefore came from the groupings that were used. These groupings essentially tell the model what our beliefs are regarding how the medications and/or CAs relate to each other, and the way in which the groupings were defined is therefore important. In an ideal situation, groups would be defined on a case-by-case basis, using expert opinion and judgements to specify relationships between different medications and CAs to achieve the most sensible groupings, in that group members should share similar properties and have the same differences in comparison to medications and CAs in other groups. The large EUROmediCAT dataset for signal detection analyses, however, means it was not practical to specify the groups on a case-by-case basis. In analyses for this thesis, there were 523 medications; it would take considerable time and effort to go through each of these individually deciding how they should be grouped. This would also require a wide range of expert opinions as the data includes a variety of different types of medication, and consensus in opinions regarding the best way to account for the differing properties and uses of the medications

would likely be difficult to achieve. Since CAs are a very heterogeneous group of conditions, with vastly differing and often unknown sets of risk factors [Oliveira and Fett-Conte, 2013], they also should not be considered as a single outcome (and this would not be possible in a signal detection dataset where there are no non-malformed controls). A pragmatic approach to grouping medications and CAs was therefore taken, by using existing hierarchies in routinely used coding systems.

For CA prevalence data, the EUROCAT coding system groups defined CAs according to the main organ system or part of the body affected by each major CA. However, chapter 3 demonstrated that EUROCAT subgroups considered to be related (in the same organ system class) were found to vary considerably in terms of their differing proportional yearly changes in prevalence. Furthermore, there are known relationships between CAs that lie within different groups of the EUROCAT hierarchy, which cannot be accounted for using this structure of coding.

For EUROmediCAT data, medications were grouped according to their ATC codes. However, relationships between characteristics captured by the ATC classification system and the actual biological or pharmacological interactions that lead to a resulting adverse drug reactions are also likely to be variable or unknown [Wang et al., 2010]. Grouping medications that have similar chemical properties therefore does not necessarily imply that they will actually have similar potential adverse effects. In practice, the ATC4 groupings might be too narrow, providing small groups and too large a total number of groups for use in statistical models taking groupings into account. On the other hand, ATC2 (or even ATC3) groupings may be considered overly broad for some types of medication. As expected for analyses in chapters 5 and 6, models considering grouping *only* by type of CA were limited, since they averaged over all medications in the data. Grouping by both CAs and medications appeared more useful in terms of the number of "high risk" signals that were identified, although the resulting effective workloads were considerably higher, and the models were computationally intensive in comparison to results obtained by a double FDR procedure.

Overall, the reliance of the performance of BHMs in this thesis on the choice of groupings of both CAs and/or medications is an important point. In particular for groupings of CA, current knowledge in this field did not allow appropriate groupings to be adopted in order to inform suitable hierarchical models for this work. Moreover, shortcomings in the available information regarding which medications are likely to be teratogenic and the absence of a gold standard list of known teratogenic medications for particular types of CA

makes the evaluation of different approaches to signal detection more difficult and uncertain.

7.3.4. Choice of thresholds used to define signals in EUROmediCAT data

An important consideration in signal detection analyses is what threshold to use to define the signals [Deshpande et al., 2010]. In this context, a "threshold" is defined as the choice of cut-off value for statistical significance of a P-value (i.e. in a Frequentist analysis such as a FDR procedure) or the percentile of the posterior distribution from an estimate in a Bayesian analysis (e.g. some lower percentile of the estimated posterior distribution for the PRR). The choice of different thresholds can change the sensitivity and number of false positive associations, hence has a potentially significant effect on any signal detection process. Other factors that may have an impact include the choice of whether to include combinations with low cell counts in the set of signals; for example, EUROmediCAT single FDR methodology only defines associations as being signals if they have at least 3 exposures.

The choice of thresholds for statistical significance when using the PRR have previously been investigated using a set of known safety signals to quantify how many would be predicted by various choices of thresholds for the PRR in SR data [Slattery et al., 2013]. This study indicated that the choice of threshold of 1 for the PRR025 was acceptable, but that an increase in the required exposure count from 3 up to a maximum of 6 may be warranted, and that this may substantially reduce the proportion of signals that are false positive associations. This would also reduce the sensitivity, however Slattery et al. [2013] suggested that any missed signals due to a loss of sensitivity in this setting would likely be detected by other means, i.e. by other aspects of their signal management processes separate from the initial signal detection analyses. For analysis of CA data, other such steps might include separate studies of specific CAs and/or medications, e.g. two of EUROmediCAT's main packages aimed to quantify the risk of CA related to four specific drug classes, and these particular medication types were therefore considered in more detail alongside signal detection analyses [de Jong-van den Berg et al., 2011]. However, unknown new potential associations can still be missed if they are in drug classes that are not considered separately or already include suspected teratogens. Other authors have also discussed statistical thresholds for use in signal detection. Ahmed et al. [2012] compared five signal detection methods, including three established methods and two methods they had previously developed based on thresholds for false positive rates. The first of these

methods is a frequentist approach that assumes a mixture model for the marginal distribution of the P-values resulting from a Fisher's exact test [Ahmed et al., 2010], and the second uses a Bayesian decision making approach using posterior probabilities of null hypotheses to rank medication–event pairs. These five methods were compared using a list of reference signals from the French national pharmacovigilance system database [Thiessard et al., 2005]. Advantages for their Bayesian method were demonstrated using the related metrics, although there was no measure of the total workload generated by each method. Another study discussed briefly the adjustment of the lower threshold for the Empirical Bayesian Geometric Mean (EBGM) and PRR estimates in order to balance the number of signals and enable a comparison of the two methods [Almenoff J. S. et al., 2006]. Berlin et al. [2012] also considered various signalling thresholds for an EBGM estimate and an OR. Their focus was on finding a threshold that maximised the number of true signals, which was to use a threshold of 2 (rather than 1) for the lower limit of the EBGM or the OR and only including those combinations with an exposure count of at least 3.

Cut-off levels used in FDR procedures

The choice of FDR cut-off threshold was discussed in Chapter 5 (see section 5.4.2). By definition, the value chosen for the FDR cut-off is the estimated proportion of the observed associations that are likely to be false positive associations. By definition for an FDR of 50%, for example, around 50% of associations are expected to be false positives, so it is essential that investigators are available to follow-up all identified associations to try and exclude the false positives. However, it is unlikely that an FDR of 50%, for example, will translate into a set of signals for which exactly half are false positives and half are true associations, in particular due to the way signals are selected in FDR procedures (i.e. after exclusions based on low frequency and PRR<1). Therefore, the FDR cut-off should be considered in terms of the resulting workload in light of available resources for investigators to follow-up potential signals at the time of each updated signal detection analysis (as a higher choice of FDR cut-off will always result in a larger workload). In practice, the use of FDR methods should therefore include frequent re-evaluation of the choice of FDR cut-off value.

Thresholds used to define signals in Bayesian hierarchical models

The thresholds used for BHMs in this thesis were a lower 2.5% confidence bound of 1 for the PRR (PRR025) and an exposure count of at least 1. These are well established thresholds that are commonly used in practice, for example by the European Medicines Agency in signal detection for their EnduraVigilance database [Alvarez et al., 2010]. Use of a

higher cut-off for the PRR (as suggested by Berlin et al. [2012]) would be overly strict if applied to the BHMs in this chapter, with a very low resulting number of signals being identified for each method; for example only 8 signals would be identified using a BHM grouped by both medications and CAs if using a cut-off of 2 for the lower limit of the 95% PCI for the PRR (see Appendix Table C2). This highlights the importance of developing signal detection methods using the specific database for which they are intended.

7.3.5. Timing of exposures in EUROmediCAT data

Strengths and limitations of EUROmediCAT data were discussed in chapter 6 (see section 6.4.3). A main strength of EUROmediCAT is in the standardised and detailed CA coding and information regarding first trimester medication use. Data cleaning was carried out prior to these analyses to ensure that only valid first trimester exposures were included. However, as with any large dataset comprising a range of data sources, it must also be acknowledged that it is impossible to capture and clean all data issues without going through all records individually (which of course is not feasible in practice). The lack of information on the exact timing of exposures and the dosage of many recorded medication exposures is an important limitation of EUROmediCAT data. No distinction can be made, for example, between a medication exposure taken at a low dose over a few days at the end of the first trimester, and exposure to the same medication at a high dose throughout the first 3 weeks of pregnancy. Furthermore, whilst the first trimester is commonly regarded as the critical period for development of most major CAs, it has also been demonstrated that different CAs have different critical periods within this time [Czeizel, 2008]. Whilst the critical period of development mostly occurs during the second and third gestational months, there are some CAs for which this period occurs after the first trimester [Scheuerle and Aylsworth, 2016]. It is therefore not possible to know whether exposures in the EUROmediCAT database actually occurred during the critical period for developments for each specific CA analysed here. This again highlights how signal detection is only part of the wider process of signal management. For any identified medication-CA signal, further consideration and investigation is required regarding the timing, duration and dosage of the particular medication exposures, and how these may relate to the known critical period for that particular CA. This further reinforces the fact that combinations of a number of different tools and data sources need to be utilised in order to obtain the timeliest, most accurate and complete measure of potential teratogenic risk, as suggested by Howard et al. [2011].

7.3.6. Evaluation of signal detection methods

Another key issue in signal detection analysis for any type of data is the lack of "gold standard" references for validation of signals when testing and comparing different methods of analysis [Almenoff June et al., 2005, Stephenson and Hauben, 2007]. Retrospective analyses applied to real data have been performed for various types of reporting datasets, which aim to measure methods in terms of their ability to identify previously known and validated signals. However, the selection of a definite set of "reference signals" is not straightforward, especially considering that the reason signal detection is being done is to find new signals and these can vary widely for different types of exposures and outcomes (i.e. they can be highly database specific). For pharmacovigilance data, the Observational Medical Outcomes Partnership (OMOP) is a research initiative which aims to recommend methods for analysis of large medication safety datasets [Stang et al., 2010, Ryan et al., 2012]. As part of this initiative, an extensively validated "gold standard" reference list of almost 400 positive and negative test cases (medication-event combinations) has been complied by the OMOP [Ryan et al., 2013]. This reference list focuses on four health outcomes (acute myocardial infarction, acute kidney injury, acute liver injury, and upper gastrointestinal bleeding) and does not include information regarding CAs. Other researchers have compiled similar lists, for example a reference standard including 10 AEs judged to be the most important outcomes across the field of pharmacovigilance was put together in order to compare signal detection methods for use with electronic healthcare records [Coloma et al., 2013]. This reference standard included 44 known positive associations and 50 highly unlikely "negative controls". Reference lists such as these have been published and are publicly available; however, they are again specific to the types of AEs included in the lists and hence are not of use for the purposes of signal detection for CA data. In other studies, retrospective real data application has also been done for some specific examples in a variety of settings, with some studies comparing the different methods [Lindquist et al., 2000, Szarfman et al., 2002, Hauben et al., 2005]. Limited "reference sets" have also been used, which have generally lacked verified true negative signals (i.e. controls) and instead focus on positive test cases only, for example see Hochberg et al. [2009] and Ahmed et al. [2012]. Ahmed et al. [2012] used three sets of up to 335 reference signals based on investigations into possible adverse medication reactions that were launched by the French pharmacovigilance system, with inclusion in a set corresponding to the level of support from pharmacovigilance data for each particular signal. The instigation of an investigation, however, does not mean that an

adverse effect was actually confirmed, and there were also no negative controls included in this reference set. All these reference lists are specific to certain types of AE, but none focus on CAs.

Another strategy has been the use of simulation studies to compare methods [Roux et al., 2005, Almenoff J. S. et al., 2006, Matsushita et al., 2007]. Simulation studies are useful for estimating and comparing methods in terms of their statistical properties and various metrics such as the quantification of false positive and negative rates. The main advantage of such an approach is that the definition of a "true" signal is definite. In particular, it is not possible to ascertain true false negatives when analysing actual databases. However, it is difficult to judge the appropriateness of any simulation model for use with real data and simulation studies need to be complemented by studies in actual reporting databases. The absolute performance of a method has been demonstrated to be highly database specific and should therefore be assessed directly on the database for which it is intended to be used in practice [Candore et al., 2015].

The lack of existing knowledge on the teratogenic effect of medicines used during pregnancy meant it was difficult to determine a "reference set" of signals on which methods used in this thesis could be evaluated (see section 4.3 for further discussion on this). Indeed, the "reference set" used to evaluate the previous EUROmediCAT signal detection analysis consisted only of eight validated medication-CA combinations [Luteijn et al., 2016]. Of these 8 combinations, one had no exposures and one had only 3 exposures in the EUROmediCAT data. For analyses in this thesis, the Australian risk categorisation system was used in an attempt to achieve a quantitative measure of each method's relative performance. Drawbacks regarding the use of this system have been discussed in previous chapters, including the lack of specificity to different CAs and the considerable proportion (35%) of ATC medications that did not have an assigned risk category. Although this categorisation database does not explicitly include "negative control" medications, it should be reasonable to expect that the "low risk" categories would be less likely to appear as signals in a signal detection analysis. This is also a problematic assumption, however, due to those signals arising as a result of confounding by indication or co-medication, i.e. where the signal represents an association likely to be due to reasons other than a teratogenic effect of the medication in question. A main weakness of the use of this risk categorisation system for signal detection analyses with CA data is that medications are classified as "high risk" in general, rather than being specific to particular CAs. However, as highlighted previously, the use of the risk categorisation system for these analyses was not to judge the

absolute strengths of a signal detection procedure, but rather to directly compare methods. The weaknesses with the use of measures based on the risk categories (e.g. “high risk” proportion and identification rate) should therefore be the same for all of the methods considered.

7.3.7. Lack of a healthy control population

Case-control studies using only malformed foetuses exposed to at least one medication cannot produce estimates of the relative odds of a CA for any medication exposure compared to healthy (i.e. non-exposed and non-malformed) controls [Prieto and Martinez-Frias, 2000]. This type of study design is therefore not able to assess whether a medication raises the risk of any CA in general in the population, and it should be emphasised that evidence of a specific increased risk of one CA in this type of study is therefore only relative to the risks associated with other CAs. On the other hand, one advantage of this type of study design is that we know, by definition of their inclusion, that each case in the data was definitely exposed to at least one medication, hence some biases associated with the use of a healthy control population (e.g. recall and recording biases) are likely to be minimised.

7.4. Potential areas for further research

7.4.1. ATC codes including multiple substances

A potential issue with the use of ATC hierarchies to group medications is that a number of substances are given different ATC codes depending on their strength or route of administration, if their therapeutic uses are clearly different [WHO Collaborating Centre for Drug Statistics Methodology, 2011]. Separate ATC codes, for example, are given for different pharmaceutical forms of a substance if applied externally to a body surface (topical) or internally administered to the circulatory system (systemic). The main therapeutic use of a medication may also differ between countries, and when a product has more than one indication there can be several options for its classification; the WHO international working group for medication statistics methodology decides upon the final classification of such substances. Using ATC3 pharmacological subgroup codes to group medications for these analyses therefore means that the same substance could be in different groups according to differing therapeutic use or local application. Although topical medications are excluded from signal detection analyses, there remain various potential routes of administration such as tablets or injections, which can be given ATC codes across different groups. Therefore, this use of grouping implies that the teratogenic risk of a

product can vary according to its usage. Furthermore, there are several "combination" products that have two or more active ingredients. These can sometimes be classified in multiple ATC3 or ATC4 levels, such that a particular substance may appear in several different ATC3 groups, depending if it is the main active ingredient or a main ingredient in combination with one or more other active ingredient. In previous EUROmediCAT signal detection analyses, substances with several potential ATC codes available were identified and their ATC5 codes replaced with the substance name [Luteijn et al., 2016]. In this thesis, however, the main aim was to assess whether grouping by ATC coding was useful and hence substances were not re-coded in this way. When these methods are implemented for use in future EUROmediCAT signal detection analyses, further consideration will need to be given as to how these substances are dealt with in terms of their grouping by ATC3 codes. For example, one approach might be to group the substances together using only one of their ATC5 codes. In the dataset for these analyses there were only 35 substances with multiple (up to ten) possible ATC codes, and it is therefore feasible that these are considered on a case-by-case basis.

7.4.2. Dealing with known teratogens in the analysis

A main aim of any signal detection study is to provide a balance between identifying the highest possible number of true signals, whilst avoiding wasting time and resources following up on false positive associations. It should also be remembered that signal detection analyses are designed to be conducted on a regular basis. This means that the number of newly generated signals will be considerably smaller for analyses after the initial run, which has already been performed for EUROmediCAT data by Luteijn et al. [2016]. Sets of signals for subsequent analyses should therefore constitute a smaller workload once discounting the known associations that do not need further follow up. It may also then be of interest to include those combinations with small cell counts in the list of potential signals, as these could allow early detection of potential teratogens. This highlights another important matter that has not been directly addressed in this thesis, which is how to deal with previously known teratogens. That is, should medications known to be teratogenic be removed from the database prior to an updated signal detection analysis? If known associations are included in the data each time they will continue to crop up as signals as long as there are new exposures occurring in the dataset. This may be the case, even for known teratogens, for example if they relate to a type of medication that is necessary for a mother's health. In the context of BHMs as used in chapter 6, it seems sensible to have kept

any known associations in the data since this may be informative to other medications within the same group for which there is less information. The double FDR procedure also presents a good argument for including known teratogens in future signal detection analyses due to the use of groupings by medications, where a strong signal in any group will ensure that this group is included in the second step of the double FDR procedure. However, it should also be noted that when doing future analyses, a medication that has already been identified as a signal may eventually "drop out" due to there being few or no new cases after they are known to be harmful (when a potential teratogenic medication is identified there may be a rapid worldwide switch to the recommendation, prescription and use of alternative medications). There are also previously reported signals that have been followed up and subsequently deemed to have been triggered by a factor other than a teratogenic effect of the medication in question; this does not, however, rule out the possibility that such a medication may actually be teratogenic, which might become evident if further exposed cases are recorded in later data updates. Some of these issues have been discussed by Lerch et al. [2015], who examined aspects of "resignalling", whereby already evaluated signals (whether reasoned to be teratogenic or not following further studies) are included in the dataset in each re-analysis. This was done according to a range of situations, including the use of different resignalling criteria and thresholds for signal detection in terms of the resulting workload. The authors concluded that some true signals will only be discovered upon resignalling.

Another related question is how to handle data from previous years in each updated analysis; should all potential data be included each time in order to increase the power, or should only the latest *X* years of data be considered in order to reduce the dimension of the data and focus on newer medications. These points also raise again the question of how frequently signal detection analyses should be performed, e.g. should it be monthly, quarterly, annually or less frequently. For each updated analysis, the updated results require careful review, placing a burden of additional work on the organisation monitoring the data. A balance is therefore required between the earliest possible detection of new signals and controlling the workload of assessing resulting signals that are false positives. Lerch et al. [2015] also evaluated the effect of varying the periodicity of analyses for their particular database and concluded that monthly signal detection analyses should be performed in order to achieve earlier detection of new potential signals. For CA data, however, there would not likely be enough new cases on a monthly basis to warrant such frequent analyses. Sufficient time needs to be allowed for new exposures to occur and for

their data to be collected and collated and EUROmediCAT registries therefore only add updated information to the central database on an annual basis. As routine signal detection analyses of EUROmediCAT data has been developed and initiated in recent years, there is currently new data available, which should be included in an updated signal detection analysis. A further analysis of EUROmediCAT data is therefore planned, using the double FDR method developed in this thesis and including additional data from 2012 onwards, as well as data from those registries that were not included in this thesis. Any new potential signals that result from this updated analysis will be fully investigated in collaboration with EUROmediCAT members and registries.

7.5. Concluding remarks

The findings of this thesis are relevant for CA surveillance programmes wishing to carry out regular statistical monitoring analyses. This applies in particular to those using EUROCAT defined subgroups for CAs and ATC coding for medications, however these conclusions also remain relevant for CA data where other types of coding are used. It is recommended that current methods of analysing each CA separately remain an appropriate for the detection of potential changes in prevalence of CAs. The double FDR procedure, however, is recommended for use in routine signal detection analyses of teratogens in CA data. The double FDR can help identify potentially teratogenic medications that may be missed by the existing single FDR procedure, and in addition may result in less false positive associations requiring detailed follow-up. This should continue to be accompanied with the follow up of any new potential signals identified as well as the dissemination of resulting information, and this will help patients make appropriate decisions based on the most up-to-date information in order to balance any risks and benefits of medication use during their pregnancy.

References

Ahmed I, Dalmasso C, Haramburu F, Thiessard F, Broet P and Tubert-Bitter P (2010). "False discovery rate estimation for frequentist pharmacovigilance signal detection methods." *Biometrics* **66**(1): 301-309.

Ahmed I, Thiessard F, Miremont-Salame G, Haramburu F, Kreft-Jais C, Begaud B and Tubert-Bitter P (2012). "Early detection of pharmacovigilance signals with automated methods based on false discovery rates: a comparative study." *Drug Saf* **35**(6): 495-506.

Akaike H (1973). "Information theory and an extension of the maximum likelihood principle." *Second International Symposium on Information Theory (Tsahkadsor, 1971)*(Akad\'emiai Kiad\'o): 267--281.

Allen VM, Armson BA, Wilson RD, Allen VM, Blight C, Gagnon A, . . . Van Aerde J (2007). "Teratogenicity associated with pre-existing and gestational diabetes." *J Obstet Gynaecol Can* **29**(11): 927-944.

Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, . . . LaCroix K (2005). "Perspectives on the Use of Data Mining in Pharmacovigilance." *Drug Safety* **28**(11): 981-1007.

Almenoff JS, LaCroix KK, Yuen NA, Fram D and DuMouchel W (2006). "Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department." *Drug Saf* **29**(10): 875-887.

Alvarez Y, Hidalgo A, Maignen F and Slattery J (2010). "Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling." *Drug Saf* **33**(6): 475-487.

Anderka M, Mitchell AA, Louik C, Werler MM, Hernández-Diaz S, Rasmussen SA and and the National Birth Defects Prevention S (2012). "Medications Used to Treat Nausea and Vomiting of Pregnancy and the Risk of Selected Birth Defects." *Birth Defects Res A Clin Mol Teratol* **94**(1): 22-30.

Andersen SL, Olsen J, Wu CS and Laurberg P (2013). "Birth defects after early pregnancy use of antithyroid drugs: a Danish nationwide study." *J Clin Endocrinol Metab* **98**(11): 4373-4381.

Australian Government Department of Health (2016). "Prescribing medicines in pregnancy database.". Accessed on 19th August, 2016, from https://www.tga.gov.au/prescribing-medicines-pregnancy-database.

Baardman ME, du Marchie Sarvaas GJ, de Walle HE, Fleurke-Rozema H, Snijders R, Ebels T, . . . Bakker MK (2014). "Impact of introduction of 20-week ultrasound scan on prevalence and fetal and neonatal outcomes in cases of selected severe congenital heart defects in The Netherlands." *Ultrasound Obstet Gynecol* **44**(1): 58-63.

Babcock GD, Talbot TO, Rogerson PA and Forand SP (2005). "Use of CUSUM and Shewhart charts to monitor regional trends of birth defect reports in New York State." *Birth Defects Res A Clin Mol Teratol* **73**(10): 669-678.

Bakker M and Jonge Ld (2014). "EUROCAT Special Report: Sources of Information on Medication Use in Pregnancy." from http://www.eurocat-network.eu/content/Special-Report-Medication-Use-In-Pregnancy.pdf.

Bate A and Evans SJ (2009). "Quantitative signal detection using spontaneous ADR reporting." *Pharmacoepidemiol Drug Saf* **18**(6): 427-436.

Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A and De Freitas RM (1998). "A Bayesian neural network method for adverse drug reaction signal generation." *Eur J Clin Pharmacol* **54**(4): 315-321.

Bates D MM, Bolker B and Walker S (2015). Fitting Linear Mixed-Effects Models using lme4. Journal of Statistical Software. **ArXiv e-print; in press**.

Begaud B, Martin K, Abouelfath A, Tubert-Bitter P, Moore N and Moride Y (2005). "An easy to use method to approximate Poisson confidence limits." *Eur J Epidemiol* **20**(3): 213-216.

Benjamini Y and Hochberg Y (1995). "Controlling the False Discovery Rate - a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society Series B-Methodological* **57**(1): 289-300.

Benjamini Y and Yekutieli D (2005). "Quantitative trait Loci analysis using the false discovery rate." *Genetics* **171**(2): 783-790.

Berlin C, Blanch C, Lewis DJ, Maladorno DD, Michel C, Petrin M, . . . Close P (2012). "Are all quantitative postmarketing signal detection methods equal? Performance characteristics of

logistic regression and Multi-item Gamma Poisson Shrinker." *Pharmacoepidemiol Drug Saf* **21**(6): 622-630.

Bernardo JM (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society. Series B (Methodological)*: 113-147.

Berry SM and Berry DA (2004). "Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model." *Biometrics* **60**(2): 418-426.

Bestwick JP, Huttly WJ, Morris JK and Wald NJ (2014). "Prevention of neural tube defects: a cross-sectional study of the uptake of folic acid supplementation in nearly half a million women." *PLoS One* **9**(2): e89354.

Blais L, Kettani F-Z, Elftouh N and Forget A (2010). "Effect of maternal asthma on the risk of specific congenital malformations: A population-based cohort study." *Birth Defects Research Part A: Clinical and Molecular Teratology* **88**(4): 216-222.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MH and White JS (2009). "Generalized linear mixed models: a practical guide for ecology and evolution." *Trends Ecol Evol* **24**(3): 127-135.

Botto LD, Lisi A, Bower C, Canfield MA, Dattani N, De Vigan C, . . . Mastroiacovo P (2006a). "Trends of selected malformations in relation to folic acid recommendations and fortification: an international assessment." *Birth Defects Res A Clin Mol Teratol* **76**(10): 693-705.

Botto LD, Robert-Gnansia E, Siffel C, Harris J, Borman B and Mastroiacovo P (2006b). "Fostering international collaboration in birth defects research and prevention: a perspective from the International Clearinghouse for Birth Defects Surveillance and Research." *Am J Public Health* **96**(5): 774-780.

Boyd PA, Haeusler M, Barisic I, Loane M, Garne E and Dolk H (2011). "Paper 1: The EUROCAT network--organization and processes." *Birth Defects Res A Clin Mol Teratol* **91 Suppl 1**: S2-15.

Boyle B, McConkey R, Garne E, Loane M, Addor MC, Bakker MK, . . . Dolk H (2013). "Trends in the prevalence, risk and pregnancy outcome of multiple births with congenital anomaly: a registry-based study in 14 European countries 1984-2007." *Bjog* **120**(6): 707-716.

Breart G (1997). "Delayed childbearing." *European Journal of Obstetrics & Gynecology and Reproductive Biology* **75**(1): 71-73.

Brook M (2011). Bayesian Hierarchical Methods for Detection of Adverse Reactions to Drugs in Large Databases using Hierarchies of Drugs and Adverse Events, London School of Hygiene and Tropical Medicine.

Brooks SP (1998). "Markov chain Monte Carlo method and its application." *Journal of the Royal Statistical Society Series D-the Statistician* **47**(1): 69-100.

Broughton NS, Graham G and Menelaus MB (1994). "The high incidence of foot deformity in patients with high-level spina bifida." *J Bone Joint Surg Br* **76**(4): 548-550.

Bruni J and Willmore LJ (1979). "Epilepsy and pregnancy." *Can J Neurol Sci* **6**(3): 345-349.

Busse R (2010). Tackling chronic disease in Europe: strategies, interventions and challenges, WHO Regional Office Europe.

Calzolari E, Barisic I, Loane M, Morris J, Wellesley D, Dolk H, . . . Garne E (2014). "Epidemiology of multiple congenital anomalies in Europe: a EUROCAT population-based registry study." *Birth Defects Res A Clin Mol Teratol* **100**(4): 270-276.

Cameron AC and Trivedi PK (2013). Regression Analysis of Count Data, Cambridge University Press.

Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S, . . . Slattery J (2015). "Comparison of Statistical Signal Detection Methods Within and Across Spontaneous Reporting Databases." *Drug Safety* **38**(6): 577-587.

Canfield MA, Honein MA, Yuskiv N, Xing J, Mai CT, Collins JS, . . . Kirby RS (2006). "National estimates and race/ethnic-specific variation of selected birth defects in the United States, 1999-2001." *Birth Defects Res A Clin Mol Teratol* **76**(11): 747-756.

Carmichael SL, Shaw GM, Laurent C, Croughan MS, Olney RS and Lammer EJ (2005). "Maternal progestin intake and risk of hypospadias." *Archives of Pediatrics & Adolescent Medicine* **159**(10): 957-962.

Casella G and George EI (1992). "Explaining the Gibbs Sampler." *American Statistician* **46**(3): 167-174.

Caster O, Norén GN, Madigan D and Bate A (2010). "Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database." *Statistical Analysis and Data Mining* **3**(4): 197-208.

Cavadino A, Prieto-Merino D, Addor MC, Arriola L, Bianchi F, Draper E, . . . Morris JK (2016). "Use of hierarchical models to analyze European trends in congenital anomaly prevalence." *Birth Defects Res A Clin Mol Teratol* **106**(6): 480-488.

Centers for Disease Control and Prevention (2010). "CDC Grand Rounds: additional opportunities to prevent neural tube defects with folic acid fortification." *MMWR Morb Mortal Wkly Rep* **59**(31): 980-984.

Chambers C (2011). "The role of teratology information services in screening for teratogenic exposures: challenges and opportunities." *Am J Med Genet C Semin Med Genet* **157c**(3): 195-200.

Charlton R, Garne E, Wang H, Klungsoyr K, Jordan S, Neville A, . . . de Jong-van den Berg L (2015). "Antiepileptic drug prescribing before, during and after pregnancy: a study in seven European regions." *Pharmacoepidemiol Drug Saf* **24**(11): 1144-1154.

Charlton RA, Klungsoyr K, Neville AJ, Jordan S, Pierini A, de Jong-van den Berg LT, . . . Garne E (2016). "Prescribing of Antidiabetic Medicines before, during and after Pregnancy: A Study in Seven European Regions." *PLoS One* **11**(5): e0155737.

Chen M, Zhu L, Chiruvolu P and Jiang Q (2015). "Evaluation of statistical methods for safety signal detection: a simulation study." *Pharm Stat* **14**(1): 11-19.

Chib S and Greenberg E (1995). "UNDERSTANDING THE METROPOLIS-HASTINGS ALGORITHM." *American Statistician* **49**(4): 327-335.

Christianson A, Howson CP and Modell B (2005). March of Dimes: global report on birth defects, the hidden toll of dying and disabled children. *White Plains*, March of Dimes Birth Defects Foundation.

Cocchi G, Gualdi S, Bower C, Halliday J, Jonsson B, Myrelid A, . . . Anneren G (2010). "International trends of Down syndrome 1993-2004: Births in relation to maternal age and terminations of pregnancies." *Birth Defects Res A Clin Mol Teratol* **88**(6): 474-479.

Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, . . . Trifiro G (2013). "A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases." *Drug Saf* **36**(1): 13-23.

Congdon P (2007). "Bayesian Statistical Modelling 2nd edition." *Biometrics* **63**(3): 976-977.

Crooks CJ, Prieto-Merino D and Evans SJ (2012). "Identifying adverse events of vaccines using a Bayesian method of medically guided information sharing." *Drug Saf* **35**(1): 61-78.

Czeizel AE (2008). "Specified critical period of different congenital abnormalities: a new approach for human teratological studies." *Congenit Anom (Kyoto)* **48**(3): 103-109.

de Jong-van den Berg L, Bakker M and Dolk H (2011). "EUROmediCAT: European surveillance of safety of medication use in pregnancy." *Pharmacoepidemiol Drug Saf* **20 (S1)**: 46–47.

de Jong J, Garne E, de Jong-van den Berg LT and Wang H (2016a). "The Risk of Specific Congenital Anomalies in Relation to Newer Antiepileptic Drugs: A Literature Review." *Drugs Real World Outcomes* **3**(2): 131-143.

de Jong J, Garne E, Wender-Ozegowska E, Morgan M, de Jong-van den Berg LT and Wang H (2016b). "Insulin analogues in pregnancy and specific congenital anomalies: a literature review." *Diabetes Metab Res Rev* **32**(4): 366-375.

de Jonge L, de Walle HE, de Jong-van den Berg LT, van Langen IM and Bakker MK (2015a). "Actual Use of Medications Prescribed During Pregnancy: A Cross-Sectional Study Using Data from a Population-Based Congenital Anomaly Registry." *Drug Saf* **38**(8): 737-747.

de Jonge L, Garne E, Gini R, Jordan SE, Klungsoyr K, Loane M, . . . Bakker MK (2015b). "Improving Information on Maternal Medication Use by Linking Prescription Data to Congenital Anomaly Registers: A EUROmediCAT Study." *Drug Saf* **38**(11): 1083-1093.

De Wals P, Tairou F, Van Allen MI, Uh SH, Lowry RB, Sibbald B, . . . Niyonsenga T (2007). "Reduction in neural-tube defects after folic acid fortification in Canada." *N Engl J Med* **357**(2): 135-142.

Deshpande G, Gogolak V and Smith SW (2010). "Data Mining in Drug Safety: Review of Published Threshold Criteria for Defining Signals of Disproportionate Reporting." *Pharmaceutical Medicine* **24**(1): 37-43.

Dolk H (2005). "EUROCAT: 25 years of European surveillance of congenital anomalies." *Arch Dis Child Fetal Neonatal Ed* **90**(5): F355-358.

Dolk H, Loane M, Garne E and European Surveillance of Congenital Anomalies Working G (2011). "Congenital heart defects in Europe: prevalence and perinatal mortality, 2000 to 2005." *Circulation* **123**(8): 841-849.

DuMouchel W (1999). "Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System." *The American Statistician* **53**(3): 177-190.

DuMouchel W and Pregibon D (2001). Empirical bayes screening for multi-item associations. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. *San Francisco, California*, ACM**:** 67-76.

Edwards IR and Biriell C (1994). "Harmonisation in pharmacovigilance." *Drug Saf* **10**(2): 93-102.

Eldridge RR, Ephross SA, Heffner CR, Tennis PS, Stender DM and White AD (1998). "Monitoring pregnancy outcomes following prenatal drug exposure through prospective pregnancy registries and passive surveillance: a pharmaceutical company commitment." *Prim Care Update Ob Gyns* **5**(4): 190-191.

Elston DA, Moss R, Boulinier T, Arrowsmith C and Lambin X (2001). "Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks." *Parasitology* **122**(Pt 5): 563-569.

Elwood JM, Little J and Elwood JH (1992). Epidemiology and control of neural tube defects. *Oxford ; New York*, Oxford University Press.

EUROCAT (2011). "EUROCAT Statistical Monitoring Report 2011." Accessed on 13th October, 2014.

EUROCAT Central Registry (2005). EUROCAT Guide 1.3: Instructions for the registration and surveillance of congenital anomalies, EUROCAT Central Registry, University of Ulster.

EUROCAT Central Registry (2009a). "Special Report, Congenital Heart Defects in Europe: 2000-2005."

EUROCAT Central Registry (2009b). Special Report: Prevention of Neural Tube Defects by Periconceptional Folic Acid Supplementation in Europe.

EUROCAT Central Registry (2012). EUROCAT Special Report: Congenital Anomalies are a Major Group of Mainly Rare Diseases.

EUROCAT Central Registry (2013). EUROCAT Guide 1.4: Instructions for the registration and surveillance of congenital anomalies, EUROCAT Central Registry, University of Ulster.

EUROCAT Central Registry (2014a). "Appendix G: EUROCAT Statistical Monitoring Report 2011." Accessed on 18th July, 2017, from http://www.eurocat-network.eu/content/Stat-Mon-Report-2011.pdf.

EUROCAT Central Registry (2014b). Special Report: Sources of Information on Medication Use in Pregnancy, EUROCAT Central Registry, University of Ulster.

EUROCAT Central Registry (2015). "EUROCAT Statistical Monitoring Report 2012.". Accessed on 15th February, 2016, from http://www.eurocat-network.eu/content/Stat-Mon-Report-2012.pdf.

EUROCAT Central Registry (2016). "What is EUROCAT?". Accessed on 23rd February, 2016, from http://www.eurocat-network.eu/aboutus/whatiseurocat/whatiseurocat.

EUROmediCAT (2011). "FP7 EUROmediCAT: WP5 - SSRIs and Anti-asthmatics." Accessed on 9th September, 2017, from http://www.euromedicat.eu/fp7euromedicat/wp5ssrisandanti-asthmatics.

EUROmediCAT (2015). Safety of Medication Use in Pregnancy: European Conference. *Poznan, Poland*.

European Medicines Agency (2012). Guideline on good pharmacovigilance practices (GVP). Module IX – Signal management.

European Medicines Agency (2016). Screening for adverse reactions in EudraVigilance.

Evans SJ, Waller PC and Davis S (2001). "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports." *Pharmacoepidemiol Drug Saf* **10**(6): 483-486.

Flour Fortification Inititative (2017). "Global Progress." Accessed on 14th July, 2017, from http://www.ffinetwork.org/global_progress/.

Fram DM, Almenoff JS and DuMouchel W (2003). Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. Proceedings of the ninth ACM SIGKDD

international conference on Knowledge discovery and data mining. *Washington, D.C.*, ACM**:** 359-368.

Garne E, Hansen AV, Morris J, Zaupper L, Addor MC, Barisic I, . . . Dolk H (2015). "Use of asthma medication during pregnancy and risk of specific congenital anomalies: A European case-malformed control study." *J Allergy Clin Immunol* **136**(6): 1496-1502.e1491-1497.

Garne E, Loane M, Dolk H, Barisic I, Addor MC, Arriola L, . . . Wiesel A (2012a). "Spectrum of congenital anomalies in pregnancies with pregestational diabetes." *Birth Defects Res A Clin Mol Teratol* **94**(3): 134-140.

Garne E, Olsen MS, Johnsen SP, Hjortdal V, Andersen HO, Nissen H, . . . Videbaek J (2012b). "How do we define congenital heart defects for scientific studies?" *Congenit Heart Dis* **7**(1): 46-49.

Garne E, Vinkel Hansen A, Morris J, Jordan S, Klungsoyr K, Engeland A, . . . Dolk H (2016). "Risk of congenital anomalies after exposure to asthma medication in the first trimester of pregnancy - a cohort linkage study." *Bjog* **123**(10): 1609-1618.

Gelman A (2006). "Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper)." *Bayesian Analysis* **1**(3): 515-533.

Gelman A and Hill J (2007). Data analysis using regression and multilevel/hierarchical models. *Cambridge; New York*, Cambridge University Press.

Gelman A, Hill J and Yajima M (2012). "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness* **5**(2): 189-211.

Gelman A and Rubin DB (1992). "Inference from Iterative Simulation Using Multiple Sequences." 457-472.

Geman S and Geman D (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *Ieee Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721-741.

Gilks WR, Best NG and Tan KKC (1995). "Adaptive Rejection Metropolis Sampling within Gibbs Sampling." *Applied Statistics-Journal of the Royal Statistical Society Series C* **44**(4): 455-472.

Giorgino FL and Egan CG (2010). "Use of isoxsuprine hydrochloride as a tocolytic agent in the treatment of preterm labour: a systematic review of previous literature." *Arzneimittelforschung* **60**(7): 415-420.

Given JE, Loane M, Luteijn JM, Morris JK, de Jong van den Berg LT, Garne E, . . . Dolk H (2016). "EUROmediCAT signal detection: an evaluation of selected congenital anomaly-medication associations." *Br J Clin Pharmacol*.

Gould AL (2003). "Practical pharmacovigilance analysis strategies." *Pharmacoepidemiol Drug Saf* **12**(7): 559-574.

Gould AL, Lystig TC, Lu Y, Fu H and Ma H (2015). "Methods and Issues to Consider for Detection of Safety Signals From Spontaneous Reporting Databases." *Therapeutic Innovation & Regulatory Science* **49**(1): 65-75.

Greenland S (2000). "Principles of multilevel modelling." *Int J Epidemiol* **29**(1): 158-167.

Greenlees R, Neville A, Addor MC, Amar E, Arriola L, Bakker M, . . . Wertelecki W (2011). "Paper 6: EUROCAT member registries: organization and activities." *Birth Defects Res A Clin Mol Teratol* **91 Suppl 1**: S51-s100.

Gregg N (1941). "Congenital cataract following German measles in the mother." *Trans Ophthalmol Soc Aust* **3**(3): 35-46.

Gunay H, Sozbilen MC, Gurbuz Y, Altinisik M and Buyukata B (2016). "Incidence and type of foot deformities in patients with spina bifida according to level of lesion." *Childs Nerv Syst* **32**(2): 315-319.

Harris BS, Bishop KC, Kemeny HR, Walker JS, Rhee E and Kuller JA (2017). "Risk Factors for Birth Defects." *Obstet Gynecol Surv* **72**(2): 123-135.

Hastings WK (1970). "Monte-Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* **57**(1): 97-&.

Hauben M, Madigan D, Gerrits CM, Walsh L and Van Puijenbroek EP (2005). "The role of data mining in pharmacovigilance." *Expert Opin Drug Saf* **4**(5): 929-948.

Hauben M and Reich L (2005). "Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: a retrospective evaluation." *J Clin Pharmacol* **45**(4): 378-384.

Hochberg AM, Hauben M, Pearson RK, O'Hara DJ, Reisinger SJ, Goldsmith DI, . . . Madigan D (2009). "An evaluation of three signal-detection algorithms using a highly inclusive reference event database." *Drug Saf* **32**(6): 509-525.

Hoffman JI and Kaplan S (2002). "The incidence of congenital heart disease." *J Am Coll Cardiol* **39**(12): 1890-1900.

Honein MA, Paulozzi LJ, Mathews TJ, Erickson JD and Wong LY (2001). "Impact of folic acid fortification of the US food supply on the occurrence of neural tube defects." *JAMA* **285**(23): 2981-2986.

Howard TB, Tassinari MS, Feibus KB and Mathis LL (2011). "Monitoring for teratogenic signals: pregnancy registries and surveillance methods." *Am J Med Genet C Semin Med Genet* **157c**(3): 209-214.

Hu JX, Zhao H and Zhou HH (2010). "False Discovery Rate Control With Groups." *J Am Stat Assoc* **105**(491): 1215-1227.

ICBDSR (2014). International Clearinghouse for Birth Defects Surveillance and Research Annual Report 2014.

Ishiguro C, Hinomura Y, Uemura K and Matsuda T (2014). "Analysis of the Factors Influencing the Spontaneous Reporting Frequency of Drug Safety Issues Addressed in the FDA's Drug Safety Communications, Using FAERS Data." *Pharmaceutical Medicine* **28**(1): 7-19.

Jentink J, Loane MA, Dolk H, Barisic I, Garne E, Morris JK and de Jong-van den Berg LTW (2010). "Valproic Acid Monotherapy in Pregnancy and Major Congenital Malformations." *New England Journal of Medicine* **362**(23): 2185-2193.

Khoshnood B, Loane M, Garne E, Addor MC, Arriola L, Bakker M, . . . Dolk H (2013). "Recent decrease in the prevalence of congenital heart defects in Europe." *J Pediatr* **162**(1): 108-113.e102.

Khoshnood B, Loane M, Walle H, Arriola L, Addor MC, Barisic I, . . . Dolk H (2015). "Long term trends in prevalence of neural tube defects in Europe: population based study." *Bmj* **351**: h5949.

Khoury MJ, Botto L, Mastroiacovo P, Skjaerven R, Castilla E and Erickson JD (1994). "Monitoring for multiple congenital anomalies: an international perspective." *Epidemiol Rev* **16**(2): 335-350.

Kirby RS and Browne ML (2016). "Editorial advances in population-based birth defects surveillance, epidemiology, and public health practice." *Birth Defects Res A Clin Mol Teratol* **106**(11): 867-868.

Kirkwood BR and Sterne J (2003). Essential medical statistics. *Oxford*, Blackwell Science.

Kruschke JK (2014). Doing Bayesian data analysis : a tutorial with R, JAGS, and stan.

Kruschke JK and Vanpaemel W (2015). Bayesian Estimation in Hierarchical Models. *The Oxford Handbook of Computational and Mathematical Psychology*. J. R. Busemeyer, Z. Wang, J. T. Townsend and A. Eidels, Oxford University Press**:** 279-299.

Lambert D (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* **34**(1): 1-14.

Lao CS (1997). "Application of CUSUM technique and beta-binomial model in monitoring adverse drug reactions." *J Biopharm Stat* **7**(2): 227-239.

Lerch M, Nowicki P, Manlik K and Wirsching G (2015). "Statistical Signal Detection as a Routine Pharmacovigilance Practice: Effects of Periodicity and Resignalling Criteria on Quality and Workload." *Drug Safety* **38**(12): 1219-1231.

Lim A, Stewart K, Konig K and George J (2011). "Systematic review of the safety of regular preventive asthma medications during pregnancy." *Ann Pharmacother* **45**(7-8): 931-945.

Lindquist M (2008). "VigiBase, the WHO Global ICSR Database System: Basic facts." *Drug Information Journal* **42**(5): 409-419.

Lindquist M and Edwards IR (2001). "The WHO Programme for International Drug Monitoring, its database, and the technical support of the Uppsala Monitoring Center." *J Rheumatol* **28**(5): 1180-1187.

Lindquist M, Stahl M, Bate A, Edwards IR and Meyboom RH (2000). "A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database." *Drug Saf* **23**(6): 533-542.

Link WA and Eaton MJ (2012). "On thinning of chains in MCMC." *Methods in Ecology and Evolution* **3**(1): 112-115.

Lisi A, Botto LD, Robert-Gnansia E, Castilla EE, Bakker MK, Bianca S, . . . Mastroiacovo P (2010). "Surveillance of adverse fetal effects of medications (SAFE-Med): findings from the international Clearinghouse of birth defects surveillance and research." *Reprod Toxicol* **29**(4): 433-442.

Lo WY and Friedman JM (2002). "Teratogenicity of recently introduced medications in human pregnancy." *Obstet Gynecol* **100**(3): 465-473.

Loane M, Dolk H and Bradbury I (2007). "Increasing prevalence of gastroschisis in Europe 1980-2002: a phenomenon restricted to younger mothers?" *Paediatr Perinat Epidemiol* **21**(4): 363-369.

Loane M, Dolk H, Garne E and Greenlees R (2011a). "Paper 3: EUROCAT data quality indicators for population-based registries of congenital anomalies." *Birth Defects Res A Clin Mol Teratol* **91 Suppl 1**: S23-30.

Loane M, Dolk H, Kelly A, Teljeur C, Greenlees R and Densem J (2011b). "Paper 4: EUROCAT statistical monitoring: identification and investigation of ten year trends of congenital anomalies in Europe." *Birth Defects Res A Clin Mol Teratol* **91 Suppl 1**: S31-43.

Loane M, Morris JK, Addor MC, Arriola L, Budd J, Doray B, . . . Dolk H (2013). "Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: impact of maternal age and prenatal screening." *Eur J Hum Genet* **21**(1): 27-33.

Lopez-Camelo JS, Orioli IM, da Graca Dutra M, Nazer-Herrera J, Rivera N, Ojeda ME, . . . Castilla EE (2005). "Reduction of birth prevalence rates of neural tube defects after folic acid fortification in Chile." *Am J Med Genet A* **135**(2): 120-125.

Lupattelli A, Spigset O, Twigg MJ, Zagorodnikova K, Mardby AC, Moretti ME, . . . Nordeng H (2014). "Medication use in pregnancy: a cross-sectional, multinational web-based study." *BMJ Open* **4**(2): e004365.

Luteijn JM, Morris JK, Garne E, Given J, de Jong-van den Berg L, Addor MC, . . . Dolk H (2016). "EUROmediCAT signal detection: a systematic method for identifying potential teratogenic medication." *Br J Clin Pharmacol*.

Maignen F, Hauben M, Hung E, Van Holle L and Dogne JM (2014). "Assessing the extent and impact of the masking effect of disproportionality analyses on two spontaneous reporting systems databases." *Pharmacoepidemiol Drug Saf* **23**(2): 195-207.

Martina R, Kay R, van Maanen R and Ridder A (2015). "The analysis of incontinence episodes and other count data in patients with overactive bladder by Poisson and negative binomial regression." *Pharm Stat* **14**(2): 151-160.

Matsushita Y, Kuroda Y, Niwa S, Sonehara S, Hamada C and Yoshimura I (2007). "Criteria revision and performance comparison of three methods of signal detection applied to the spontaneous reporting database of a pharmaceutical manufacturer." *Drug Saf* **30**(8): 715-726.

McBride WG (1961). "Thalidomide and congenital abnormalities." *The Lancet* **278**(7216): 1358.

Mehrotra DV and Adewale AJ (2012). "Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals." *Stat Med* **31**(18): 1918-1930.

Mehrotra DV and Heyse JF (2004). "Use of the false discovery rate for evaluating clinical safety data." *Stat Methods Med Res* **13**(3): 227-238.

Meijer WM, Cornel MC, Dolk H, de Walle HE, Armstrong NC and de Jong-van den Berg LT (2006). "The potential of the European network of congenital anomaly registers (EUROCAT) for drug safety surveillance: a descriptive study." *Pharmacoepidemiol Drug Saf* **15**(9): 675-682.

Mills JL (2010). "Malformations in Infants of Diabetic Mothers." *Birth Defects Res A Clin Mol Teratol* **88**(10): 769-778.

Mitchell AA (2016). "Research challenges for drug-induced birth defects." *Clin Pharmacol Ther* **100**(1): 26-28.

Mitchell AA, Gilboa SM, Werler MM, Kelley KE, Louik C and Hernandez-Diaz S (2011). "Medication use during pregnancy, with particular focus on prescription drugs: 1976-2008." *Am J Obstet Gynecol* **205**(1): 51.e51-58.

Morgan M, De Jong-van den Berg LT and Jordan S (2011). "Drug safety in pregnancy--monitoring congenital anomalies." *J Nurs Manag* **19**(3): 305-310.

Mosley JF, Smith LL and Dezan MD (2015). "An overview of upcoming changes in pregnancy and lactation labeling information." *Pharmacy Practice* **13**(2): 605.

MRC Vitamin Study Research Group (1991). "Prevention of neural tube defects: results of the Medical Research Council Vitamin Study." *Lancet* **338**(8760): 131-137.

Mullahy J (1986). "SPECIFICATION AND TESTING OF SOME MODIFIED COUNT DATA MODELS." *Journal of Econometrics* **33**(3): 341-365.

Nau H (2001). "Teratogenicity of isotretinoin revisited: species variation and the role of all-trans-retinoic acid." *J Am Acad Dermatol* **45**(5): S183-187.

Neal RM (2003). "Slice sampling." *Annals of Statistics* **31**(3): 705-741.

Newson R (2003). "Multiple-test procedures and smile plots." *Stata Journal* **3**(2): 109-132.

Ntzoufras I (2009). Bayesian Modeling Using WinBUGS. *New Jersey*, John Wiley & Sons, Inc.

Oliveira CI and Fett-Conte AC (2013). "Birth defects: Risk factors and consequences." *J Pediatr Genet* **2**(2): 85-90.

Pariente A, Avillach P, Salvo F, Thiessard F, Miremont-Salame G, Fourrier-Reglat A, . . . Moore N (2012). "Effect of competition bias in safety signal generation: analysis of a research database of spontaneous reports in France." *Drug Saf* **35**(10): 855-864.

Parker SE, Mai CT, Canfield MA, Rickard R, Wang Y, Meyer RE, . . . Correa A (2010). "Updated National Birth Prevalence estimates for selected birth defects in the United States, 2004-2006." *Birth Defects Res A Clin Mol Teratol* **88**(12): 1008-1016.

Petersen I, Collings SL, McCrea RL, Nazareth I, Osborn DP, Cowen PJ and Sammon CJ (2017). "Antiepileptic drugs prescribed in pregnancy and prevalence of major congenital malformations: comparative prevalence studies." *Clin Epidemiol* **9**: 95-103.

Plummer M (2002). "Discussion of the paper by Spiegelhalter et al." *Journal of the Royal Statistical Society Series B* **64**: 620.

Plummer M (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. 3rd International Workshop on Distributed Statistical Computing (DSC 2003). F. L. Kurt Hornik, amp and Z. Achim. *Technische Universität Wien, Vienna, Austria*.

Plummer M, Best N, A Cowles K and A Vines K (2003). CODA: convergence diagnosis and output analysis for MCMC. R News. **6:** 7-11.

Poletta FA, Gili JA and Castilla EE (2014). "Latin American Collaborative Study of Congenital Malformations (ECLAMC): a model for health collaborative studies." *Public Health Genomics* **17**(2): 61-67.

Prieto-Merino D, Quartey G, Wang J and Kim J (2011). "Why a Bayesian approach to safety analysis in pharmacovigilance is important." *Pharm Stat* **10**(6): 554-559.

Prieto L and Martinez-Frias ML (2000). "Case-control studies using only malformed infants who were prenatally exposed to drugs. What do the results mean?" *Teratology* **62**(1): 5-9.

Raiffa H and Schlaifer R (1961). Applied statistical decision theory. *Boston*, Division of Research, Graduate School of Business Adminitration, Harvard University.

Ramoz LL and Patel-Shori NM (2014). "Recent changes in pregnancy and lactation labeling: retirement of risk categories." *Pharmacotherapy* **34**(4): 389-395.

Ross GJS and Preece DA (1985). "The Negative Binomial Distribution." *Journal of the Royal Statistical Society. Series D (The Statistician)* **34**(3): 323-335.

Rothman KJ (1990). "No adjustments are needed for multiple comparisons." *Epidemiology* **1**(1): 43-46.

Roux E, Thiessard F, Fourrier A, Begaud B and Tubert-Bitter P (2005). "Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance." *IEEE Trans Inf Technol Biomed* **9**(4): 518-527.

Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA and Hartzema AG (2012). "Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership." *Stat Med* **31**(30): 4401-4415.

Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S and Hartzema AG (2013). "Defining a reference set to support methodological research in drug safety." *Drug Saf* **36 Suppl 1**: S33-47.

Sachdeva P, Patel BG and Patel BK (2009). "Drug Use in Pregnancy; a Point to Ponder!" *Indian Journal of Pharmaceutical Sciences* **71**(1): 1-7.

Saefken B, Kneib T, van Waveren C-S and Greven S (2014). "A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models." 201-225.

Saefken B and Ruegamer D (2014). {cAIC4}: Conditional {Akaike} information criterion for lme4. R package version 0.1.

Sakaeda T, Tamon A, Kadoyama K and Okuno Y (2013). "Data mining of the public version of the FDA Adverse Event Reporting System." *Int J Med Sci* **10**(7): 796-803.

Sannerstedt R, Lundborg P, Danielsson BR, Kihlstrom I, Alvan G, Prame B and Ridley E (1996). "Drugs during pregnancy: an issue of risk classification and information to prescribers." *Drug Saf* **14**(2): 69-77.

Savitz DA and Olshan AF (1995). "Multiple comparisons and related issues in the interpretation of epidemiologic data." *Am J Epidemiol* **142**(9): 904-908.

Savva GM, Walker K and Morris JK (2010). "The maternal age-specific live birth prevalence of trisomies 13 and 18 compared to trisomy 21 (Down syndrome)." *Prenat Diagn* **30**(1): 57-64.

Schaefer C, Hannemann D and Meister R (2005). "Post-marketing surveillance system for drugs in pregnancy--15 years experience of ENTIS." *Reprod Toxicol* **20**(3): 331-343.

Scheuerle AE and Aylsworth AS (2016). "Birth defects and neonatal morbidity caused by teratogen exposure after the embryonic period." *Birth Defects Res A Clin Mol Teratol* **106**(11): 935-939.

Schonfeld T, Schmid KK, Brown JS, Amoura NJ and Gordon B (2013). "A pregnancy testing policy for women enrolled in clinical trials." *Irb* **35**(6): 9-15.

Sedgh G, Singh S and Hussain R (2014). "Intended and Unintended Pregnancies Worldwide in 2012 and Recent Trends." *Studies in family planning* **45**(3): 301-314.

Sharrar RG and Dieck GS (2013). "Monitoring product safety in the postmarketing environment." *Ther Adv Drug Saf* **4**(5): 211-219.

Shields KE and Lyerly AD (2013). "Exclusion of pregnant women from industry-sponsored clinical trials." *Obstet Gynecol* **122**(5): 1077-1081.

Slattery J, Alvarez Y and Hidalgo A (2013). "Choosing thresholds for statistical signal detection with the proportional reporting ratio." *Drug Saf* **36**(8): 687-692.

Spiegelhalter DJ, Best NG, Carlin BP and Van Der Linde A (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4): 583-639.

Stan Development Team (2015a). RStan: the R interface to Stan, Version 2.7.0.

Stan Development Team (2015b). Stan: A C++ Library for Probability and Sampling, Version 2.8.0.

Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, . . . Woodcock J (2010). "Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership." *Ann Intern Med* **153**(9): 600-606.

StataCorp (2011). Stata Statistical Software: Release 12., College Station, TX: StataCorp LP.

Stephenson WP and Hauben M (2007). "Data mining for signals in spontaneous reporting databases: proceed with caution." *Pharmacoepidemiol Drug Saf* **16**(4): 359-365.

Suling M and Pigeot I (2012). "Signal detection and monitoring based on longitudinal healthcare data." *Pharmaceutics* **4**(4): 607-640.

Szarfman A, Machado SG and O'Neill RT (2002). "Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database." *Drug Saf* **25**(6): 381-392.

Szarfman A, Tonning JM and Doraiswamy PM (2004). "Pharmacovigilance in the 21st century: new systematic tools for an old problem." *Pharmacotherapy* **24**(9): 1099-1104.

Tatonetti NP, Ye PP, Daneshjou R and Altman RB (2012). "Data-driven prediction of drug effects and interactions." *Sci Transl Med* **4**(125): 125ra131.

Thiessard F, Roux E, Miremont-Salame G, Fourrier-Reglat A, Haramburu F, Tubert-Bitter P and Begaud B (2005). "Trends in spontaneous adverse drug reaction reports to the French pharmacovigilance system (1986-2001)." *Drug Saf* **28**(8): 731-740.

Tierney L (1994). "Markov Chains for Exploring Posterior Distributions." *The Annals of Statistics* **22**(4): 1701-1728.

Tomson T, Battino D, Bonizzoni E, Craig J, Lindhout D, Sabers A, . . . Vajda F (2011). "Dose-dependent risk of malformations with antiepileptic drugs: an analysis of data from the EURAP epilepsy and pregnancy registry." *Lancet Neurol* **10**(7): 609-617.

Tubert P, Bégaud B, Péré J-C, Haramburu F and Lellouch J (1992). "Power and weakness of spontaneous reporting: A probabilistic approach." *Journal of Clinical Epidemiology* **45**(3): 283-286.

US Food and Drug Administration (2014). "Promoting your health. Pregnancy and Lactation Labeling Final Rule.". Accessed on 02 September, 2016, from http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm425317.htm.

US Food and Drug Administration (2016). "FDA Adverse Event Reporting System (FAERS). ." Accessed on 7th August, 2017, from https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/.

Vaida F and Blanchard S (2005). "Conditional Akaike information for mixed-effects models." *Biometrika* **92**(2): 351-370.

van der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJ and Roos-Hesselink JW (2011). "Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis." *J Am Coll Cardiol* **58**(21): 2241-2247.

van Gelder MM, de Jong-van den Berg LT and Roeleveld N (2014). "Drugs associated with teratogenic mechanisms. Part II: a literature review of the evidence on human risks." *Hum Reprod* **29**(1): 168-183.

van Gelder MM, van Rooij IA, Miller RK, Zielhuis GA, de Jong-van den Berg LT and Roeleveld N (2010). "Teratogenic mechanisms of medical drugs." *Hum Reprod Update* **16**(4): 378-394.

van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R and Egberts AC (2002). "A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions." *Pharmacoepidemiol Drug Saf* **11**(1): 3-10.

Vauzelle C, Beghin D, Cournot M-P and Elefant E (2013). "Birth defects after exposure to misoprostol in the first trimester of pregnancy: Prospective follow-up study." *Reproductive Toxicology* **36**: 98-103.

Ver Hoef JM and Boveng PL (2007). "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?" *Ecology* **88**(11): 2766-2772.

Veroniki AA, Cogo E, Rios P, Straus SE, Finkelstein Y, Kealey R, . . . Tricco AC (2017). "Comparative safety of anti-epileptic drugs during pregnancy: a systematic review and network meta-analysis of congenital malformations and prenatal outcomes." *BMC Med* **15**(1): 95.

Waller P, van Puijenbroek E, Egberts A and Evans S (2004). "The reporting odds ratio versus the proportional reporting ratio: 'deuce'." *Pharmacoepidemiol Drug Saf* **13**(8): 525-526; discussion 527-528.

Wang HW, Hochberg AM, Pearson RK and Hauben M (2010). "An experimental investigation of masking in the US FDA adverse event reporting system database." *Drug Saf* **33**(12): 1117-1133.

Wemakor A, Casson K, Garne E, Bakker M, Addor MC, Arriola L, . . . Dolk H (2015). "Selective serotonin reuptake inhibitor antidepressant use in first trimester pregnancy and risk of specific congenital anomalies: a European register-based study." *Eur J Epidemiol* **30**(11): 1187-1198.

WHO Collaborating Centre for Drug Statistics Methodology (2011, 2011-03-25). "ATC Structure and principles." Accessed on 21 April, 2016, from http://www.whocc.no/atc/structure_and_principles/.

WHO Collaborating Centre for Drug Statistics Methodology (2015, 18 December 2015). "ATC alterations from 1982-2016." Accessed on 11 August, 2016, from http://www.whocc.no/atc_ddd_alterations__cumulative/atc_alterations/.

WHO/CDC/ICBDSR (2014). Birth defects surveillance: a manual for programme managers. *Geneva*, World Health Organization.

Wilson JG (1979). "The evolution of teratological testing." *Teratology* **20**(2): 205-211.

Wilson RD, Audibert F, Brock JA, Carroll J, Cartier L, Gagnon A, . . . Van den Hof M (2015). "Pre-conception Folic Acid and Multivitamin Supplementation for the Primary and Secondary Prevention of Neural Tube Defects and Other Folic Acid-Sensitive Congenital Anomalies. SOGC Clinical Practice Guideline no. 324, May 2015." *J Obstet Gynaecol Can* **37**(6): 534-552.

World Health Organization (2016). "Fact sheet N° 370 Congenital anomalies." *World Health Organization Fact sheets*. Accessed on June 20, 2017, from http://www.who.int/mediacentre/factsheets/fs370/en/.

Xia HA, Ma H and Carlin BP (2011). "Bayesian hierarchical modeling for detecting safety signals in clinical trials." *J Biopharm Stat* **21**(5): 1006-1029.

Yakoob MY, Bateman BT, Ho E, Hernandez-Diaz S, Franklin JM, Goodman JE and Hoban RA (2013). "THE RISK OF CONGENITAL MALFORMATIONS ASSOCIATED WITH EXPOSURE TO BETA-BLOCKERS EARLY IN PREGNANCY: A META-ANALYSIS." *Hypertension* **62**(2): 375-381.

Zabihi S and Loeken MR (2010). "Understanding diabetic teratogenesis: where are we now and where are we going?" *Birth Defects Res A Clin Mol Teratol* **88**(10): 779-790.

Zaganjor I, Sekkarie A, Tsang BL, Williams J, Razzaghi H, Mulinare J, . . . Rosenthal J (2016). "Describing the Prevalence of Neural Tube Defects Worldwide: A Systematic Literature Review." *PLoS One* **11**(4): e0151586.

Zaqout M, Aslem E, Abuqamar M, Abughazza O, Panzer J and De Wolf D (2015). "The Impact of Oral Intake of Dydrogesterone on Fetal Heart Development During Early Pregnancy." *Pediatr Cardiol* **36**(7): 1483-1488.

Zeinoun Z, Seifert H and Verstraeten T (2009). "Quantitative signal detection for vaccines: effects of stratification, background and masking on GlaxoSmithKline's spontaneous reports database." *Hum Vaccin* **5**(9): 599-607.

Appendices

Appendix A: Supplementary material for Chapter 3

Table A1. Coding of EUROCAT subgroups included in analysis of changes in prevalence of CAs (taken from EUROCAT Guide 1.3 Table 8.2).

Anomaly	Description Of Anomaly	ICD10 code	Comments
All Congenital Anomalies		Q*, D215, D821, D1810, P350, P351, P371	Exclude all minor anomalies
Nervous system		**Q00, Q01, Q02, Q03, Q04, Q05, Q06, Q07**	
Neural tube defects	Neural tube defects include anencephalus, encephalocele, spina bifida and iniencephalus	Q00, Q01, Q05	
Anencephalus and similar	Total or partial absence of brain tissue and the cranial vault. The face and eyes are present. (incompatible with life)	Q00	
Encephalocele	Cystic expansion of meninges and brain tissue outside the cranium. Covered by normal or atrophic skin.	Q01	exclude if associated with anencephalus
Spina Bifida	Midline defect of the osseous spine usually affecting the posterior arches resulting in a herniation or exposure of the spinal cord and/or meninges	Q05	exclude if associated with anencephalus, or encephalocele
Congenital heart defects		**Q20-Q26**	**exclude patent ductus arteriosus (PDA) in preterm/LBW babies (<37 weeks) - ICD9: 7470; ICD10: Q250**

Severe CHD §	Severe congenital heart defects have higher perinatal mortality and TOPFA rates. Most livebirths require surgery for survival. It includes: single ventricle, tricuspid atresia, Ebstein's anomaly, hypoplastic left heart, hypoplastic right heart, common arterial truncus, transposition of great vessels, atrioventricular septal defects, tetralogy of fallot, pulmonary valve atresia, aortic valve atresia/stenosis, coarctation of aorta and total anomalous pulmonary venous return.	Q200, Q203-Q204, Q212-Q213, Q220, Q224-Q226,Q230, Q234, Q251, Q262
Common arterial truncus	Presence of a large single arterial vessel at the base of the heart (from which the aortic arch, pulmonary and coronary arteries originate), always accompanied by a large subvulvar septal defect.	Q200
Transposition of great vessels	Total separation of circulation with the aorta arising from the right ventricle and the pulmonary artery from the left ventricle	Q203
Single ventricle	Only one complete ventricle with an inlet valve and an outlet portion even though the outlet valve is atretic	Q204
Ventricular septal defect	Defect in the ventricular septum	Q210
Atrial septal defect	Defect in the atrial septum	Q211
Atrioventricular septal defect	Central defect of the cardiac septa and a common atrioventricular valve, includes primum ASD defects	Q212
Tetralogy of Fallot	VSD close to the aortic valves, infundibular and pulmonary valve stenosis and over-riding aorta across the VSD	Q213
Tricuspid atresia and stenosis	Obstruction of the tricuspid valve and hypoplasia of the right ventricle	Q224
Ebstein's anomaly	Tricuspid valve displaced with large right atrium and	Q225

	small right ventricle	
Pulmonary valve stenosis	Obstruction or narrowing of the pulmonary valves which may impair blood flow through the valves	Q221
Pulmonary valve atresia	Lack of patency or failure of formation altogether of the pulmonary valve, resulting in obstruction of the blood flow from the right ventricle to the pulmonary artery	Q220
Aortic valve atresia/stenosis §	Occlusion of aortic valve or stenosis of varying degree, often associated with bicuspid valves	Q230
Hypoplastic left heart	Hypoplasia of the left ventricle, outflow tract and ascending aorta resulting from an obstructive lesion of the left side of the heart	Q234
Hypoplastic right heart §	Hypoplasia of the right ventricle, always associated with other cardiac malformations	Q226
Coarctation of aorta	Constriction in the region of aorta where the ductus joins aorta	Q251
Total anomalous pulmonary venous return	All four pulmonary veins drain to right atrium or one of the venous tributaries	Q262
PDA as only CHD in term infants (>=37 weeks)		Q250
Digestive system		**Q38-Q39, Q402-Q409, Q41-Q45 exclude Q381, Q382, Q3850, Q430, Q4320, Q4381, Q4382**
Oesophageal atresia with or without tracheo-oesophageal fistula	Occlusion or narrowing of the oesophagus with or without tracheo-oesophagael fistula	Q390-Q391

Duodenal atresia or stenosis	Occlusion or narrowing of duodenum	Q410	exclude if also annular pancreas (Q451, 75172)
Atresia or stenosis of other parts of small intestine	Occlusion or narrowing of other parts of small intestine	Q411-Q418	
Ano-rectal atresia and stenosis	Imperforate anus or absence or narrowing of the communication canal between the rectum and anus with or without fistula to neighbouring organs	Q420-Q423	
Hirschsprung's disease	Absence of the parasympatic ganglion nerve cells (aganglionosis) of the wall of the colon or rectum. May result in cong megacolon	Q431	
Atresia of bile ducts	Congenital absence of the lumen of the extrahepatic bile ducts	Q442	
Annular pancreas	pancreas surrounds the duodenum causing stenosis	Q451	
Diaphragmatic hernia	Defect in the diaphragm with protrusion of abdominal content into the thoracic cavity. Various degree of lung hypoplasia on the affected side	Q790	
Chromosomal		**Q90-Q93, Q96-Q99 exclude Q936**	**microdeletions excluded**
Down Syndrome/trisomy 21	karyotype 47,xx +21 or 47,xy +21 and translocations/mosaicism	Q90	
Patau syndrome/trisomy 13	karyotype 47,xx +13 or 47,xy +13 and translocations/mosaicism	Q914-Q917	
Edwards syndrome/trisomy 18	karyotype 47,xx +18 or 47,xy +18 and translocations/mosaicism	Q910-Q913	
Turner syndrome	karyotype 45,x or structural anomalies of X chromosome	Q96	
Klinefelter syndrome	karyotype 47,xxy or additional X-chromosomes	Q980-Q984	

A1. Bonferroni adjustment to confidence intervals for multiple testing within groups of congenital anomalies

Adjustments to 95% CIs for estimated trends in frequentist individual models were done using a Bonferroni correction to adjust for the number of Congenital Anomalies (CAs) in each analysis. This reflects the fact that estimates in individual models are being compared to Bayesian hierarchical models where multiple tests are not done since all the CAs are included in one model.

In practise this means that instead of a 95% confidence level, a $\left(1-\frac{\alpha}{k}\right)\%$ interval was used, where k is the number of tests in each group and $\alpha = 5\%$. If the estimated yearly change in prevalence β is assumed to have an approximate Normal distribution with zero mean and standard error $SE(\beta)$, then a $(1-\alpha)\%$ CI for β is constructed as

$$(1-\alpha)\%\ CI = \beta \pm z_{\alpha/2} * SE(\beta)$$

Here $z_{\alpha/2}$ is the critical value of the standard Normal distribution. For a 95% CI, $z_{\alpha/2} = z_{0.05} = 1.96$, hence

$$95\%\ CI = \beta \pm 1.96 * SE(\beta)$$

Adjusted values for $z_{0.025}$ were used to calculate the adjusted 95% CIs when taking into account multiple testing within groups of CAs in individual Poisson models

$$95\%\ CIadj \quad = \quad \beta \pm z_{0.025/k} * SE(\beta) \quad = \quad \beta \pm zadj_{0.025} * SE(\beta)$$

Table A2. Critical values used to adjust 95% CIs for multiple testing in individual models for groups of CAs in chapter 3.

Type of CA	**Number of CAs** k	**Adjusted confidence level** $\left(1-\frac{0.05}{k}\right)\%$	**Adjusted $z_{0.025}$ value for calculation of adjusted CIs** $(zadj_{0.025})$
Neural tube defects	3	98.33	2.34
Chromosomal	5	99.00	2.58
Digestive system	16	99.39	2.95
Congenital heart defects	8	99.38	2.74

Table A3. R and JAGS code for models described in section 3.5.3.

Model	Example R/JAGS code
1	`glmer(cases ~ yr + (1|obs.eff), offset=log(totalb), data=data, family=poisson(link="log"))`
2	`glmer(cases ~ yr + (1|centre) + (1|obs.eff), offset=log(totalb), data=data, family=poisson(link="log"))`
3	`glmer(cases ~ yr + (yr|CA) + (1|obs.eff), offset=log(totalb), data=data, family=poisson(link="log"))`
4	`glmer(cases ~ yr + (yr|CA) + (1|centre) + (1|obs.eff), offset=log(totalb), data=data, family=poisson(link="log"))`

5	`model {`
	`  for (i in 1:n.obs) {`
	`  y[i] ~ dpois(lambda[i])`
	`  log(lambda[i]) <- mu[i]`
	`  mu[i] <-  offs[i] + u0[CA[i]] + u1[CA[i]]*x[i] + epsilon[i]`
	`  epsilon[i] ~ dnorm(0, tau.e)`
	`  }`
	
	`  tau.e <- pow(sigma.e,-2)`
	`  sigma.e ~ dunif(0, 10)`
	
	`  for (k in 1:n.CA) {`
	`  u0[k] ~ dnorm(mu.u0, tau.u0)`
	`  u1[k] ~ dnorm(mu.u1, tau.u1)`
	`  }`
	
	`  mu.u0 ~ dnorm(0, 0.001)`
	`  tau.u0 <- pow(sigma.u0,-2)`
	`  sigma.u0 ~ dunif(0, 10)`
	`  mu.u1 ~ dnorm(0, 0.001)`
	`  tau.u1 <- pow(sigma.u1,-2)`
	`  sigma.u1 ~ dunif(0, 10)`
	`  }`

6A

```
model {
    for (i in 1:n.obs) {
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- mu[i]
      mu[i]<-offs[i] + u[reg[i]] + u0[CA[i]]+ u1[CA[i]]*x[i] +
epsilon[i]
      epsilon[i] ~ dnorm(0, tau.e)
    }
    tau.e <- pow(sigma.e,-2)
    sigma.e ~ dunif(0, 10)

    for (j in 1:n.reg) {
      u[j] ~ dnorm(mu.u, tau.u)
    }
    for (k in 1:n.CA) {
      u0[k] ~ dnorm(mu.u0, tau.u0)
      u1[k] ~ dnorm(mu.u1, tau.u1)
    }

    mu.u ~ dnorm(0, 0.001)
    tau.u <- pow(sigma.u,-2)
    sigma.u ~ dunif(0, 10)
    mu.u0 ~ dnorm(0, 0.001)
    tau.u0 <- pow(sigma.u0,-2)
    sigma.u0 ~ dunif(0, 10)
    mu.u1 ~ dnorm(0, 0.001)
    tau.u1 <- pow(sigma.u1,-2)
    sigma.u1 ~ dunif(0, 10)
  }
```

6B

```
model {
    for (i in 1:n.obs) {
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- mu[i]
      mu[i]<-offs[i] + u0[reg[i]] + u1[CA[i]]*x[i] + epsilon[i]
      epsilon[i] ~ dnorm(0, tau.e)
    }
    tau.e <- pow(sigma.e,-2)  # Priors
    sigma.e ~ dunif(0, 10)

    for (j in 1:n.reg) {
      u0[j] ~ dnorm(mu.u0[CA[j]], tau.u0[CA[j]])
    }
    for (k in 1:n.CA) {
      mu.u0[k] ~ dnorm(mu.jk, tau.jk)
      tau.u0[k] <- pow(sigma.u0[k], -2)
      sigma.u0[k] ~ dunif(0, 10)
      u1[k] ~ dnorm(mu.u1, tau.u1)
    }
    mu.jk  ~ dnorm(0, 0.001)
    tau.jk <- pow(sigma.jk, -2)
    sigma.jk ~ dunif(0, 10)
    mu.u1 ~ dnorm(0, 0.001)
    tau.u1 <- pow(sigma.b,-2)
    sigma.u1 ~ dunif(0, 10)
  }
```

Table A4. Summary of notation for models in chapter 3.

Model	Model notation
Model 1: Individual models	$y_t \sim Poisson(\lambda_t)$ $\log(\lambda_t) = log(p_t) + \beta_0 + \beta_1 x_t + \varepsilon_t$ $\varepsilon_t \sim normal(0, \sigma_\varepsilon^2)$
Model 2: Individual models + registry	$y_{tj} \sim Poisson(\lambda_{tj})$ $log(\lambda_{tj}) = log(p_{tj}) + (\beta_0 + u_j) + \beta_1 x_{tj} + \varepsilon_{tj}$ $u_j \sim normal(0, \sigma_j^2)$ $\varepsilon_{tj} \sim normal(0, \sigma_\varepsilon^2)$
Model 3: Frequentist hierarchical model	$y_{tk} \sim Poisson(\lambda_{tk})$ $log(\lambda_{tk}) = log(p_{tk}) + (\beta_0 + u_{0k}) + (\beta_1 + u_{1k}) x_{tk} + \varepsilon_{tk}$ $u_{0k} \sim normal(0, \sigma_{0k}^2)$ $u_{1k} \sim normal(0, \sigma_{1k}^2)$ $\varepsilon_{tk} \sim normal(0, \sigma_\varepsilon^2)$
Model 4: Frequentist hierarchical model + registry	$y_{tjk} \sim Poisson(\lambda_{tjk})$ $log(\lambda_{tjk}) = log(p_{tjk}) + (\beta_0 + u_{0k} + u_j) + (\beta_1 + u_{1k}) x_{tjk} + \varepsilon_{tjk}$ $u_{0k} \sim normal(0, \sigma_{0k}^2)$ $u_{1k} \sim normal(0, \sigma_{1k}^2)$ $u_j \sim normal(0, \sigma_j^2)$ $\varepsilon_{tjk} \sim normal(0, \sigma_\varepsilon^2)$
Model 5: Bayesian hierarchical model	$y_{tk} \sim Poisson(\lambda_{tk})$ $log(\lambda_{tk}) = log(p_{tk}) + u_{0k} + u_{1k} x_{tk} + \varepsilon_{tk}$ $u_{0k} \sim normal(\mu_{u0}, \tau_{u0})$ $u_{1k} \sim normal(\mu_{u1}, \tau_{u1})$ $\varepsilon_{tk} \sim normal(0, \tau_\varepsilon)$ $\mu_{u0},\ \mu_{u1} \sim normal(0, 0.001)$ $\tau_{u0} = \frac{1}{\sigma_{u0}^2}, \tau_{u1} = \frac{1}{\sigma_{u1}^2}, \tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2}$ $\sigma_{u0}, \sigma_{u1}, \sigma_{\varepsilon k} \sim uniform(0, 10)$

Model 6A: Bayesian hierarchical model + registry (A)	$$y_{tjk} \sim Poisson(\lambda_{tjk})$$ $$log(\lambda_{tjk}) = log(p_{tjk}) + u_j + u_{0k} + u_{1k}x_{tjk} + \varepsilon_{tjk}$$ $$u_{0k} \sim normal(\mu_{u0}, \tau_{u0})$$ $$u_{1k} \sim normal(\mu_{u1}, \tau_{u1})$$ $$u_j \sim normal(\mu_j, \tau_j)$$ $$\varepsilon_{tk} \sim normal(0, \tau_\varepsilon)$$ $$\mu_{u0},\ \mu_{u1} \sim normal(0, 0.001)$$ $$\tau_{u0} = \frac{1}{\sigma_{u0}^2}, \tau_{u1} = \frac{1}{\sigma_{u1}^2}, \tau_j = \frac{1}{\sigma_j^2},\ \tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2}$$ $$\sigma_{u0}, \sigma_{u1}, \sigma_j, \sigma_{\varepsilon k} \sim uniform(0, 10)$$
Model 6B: Bayesian hierarchical model + registry (B)	$$y_{tjk} \sim Poisson(\lambda_{tjk})$$ $$log(\lambda_{tjk}) = log(p_{tjk}) + u_{0jk} + u_{1k}x_{tjk} + \varepsilon_{tjk}$$ $$u_{0jk} \sim normal(\mu_{u0}, \tau_{u0})$$ $$u_{1k} \sim normal(\mu_{u1}, \tau_{u1})$$ $$\varepsilon_{tk} \sim normal(0, \tau_\varepsilon)$$ $$\mu_{u0},\ \mu_{u1} \sim normal(0, 0.001)$$ $$\tau_{u0} = \frac{1}{\sigma_{u0}^2}, \tau_{u1} = \frac{1}{\sigma_{u1}^2}, \tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2}$$ $$\sigma_{u0}, \sigma_{u1}, \sigma_j, \sigma_{\varepsilon k} \sim uniform(0, 10)$$

A2. Example script to run a Bayesian hierarchical model in R and JAGS

```
## MODEL 5. Bayesian Poisson model grouping CAs, pooling over registry

# model specification
cat("model {

    # Likelihood
    for (i in 1:n.obs) {
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- mu[i]
      mu[i] <-  offs[i] + u0[CA[i]] + u1[CA[i]]*x[i] + epsilon[i]
      epsilon[i] ~ dnorm(0, tau.e)
    }

    # Priors
    tau.e <- pow(sigma.e,-2)
    sigma.e ~ dunif(0, 100)

    for (k in 1:n.CA) {
      u0[k] ~ dnorm(mu.u0, tau.u0)
      u1[k] ~ dnorm(mu.u1, tau.u1)
    }

    # hyperpriors
    mu.u0 ~ dnorm(0, 0.001)
    tau.u0 <- pow(sigma.u0,-2)
    sigma.u0 ~ dunif(0, 10)
    mu.u1 ~ dnorm(0, 0.001)
    tau.u1 <- pow(sigma.u1,-2)
    sigma.u1 ~ dunif(0, 10)
 }", file="m5.txt")

# specify the total number of registries and CA subgroups
n.reg=length(levels(usedat5$centre))
n.CA=length(levels(usedat5$CA))

# set parameters to be monitored
params5 <- c("u0","mu.u0","sigma.u0","u1","mu.u1","sigma.u1","sigma.e")

# set initial values for each chain using random numbers from the
normal and uniform distributions
inits.m5 <- function(chain) return(switch(chain,
  "1"=list(u0=10*rnorm(n.CA), mu.u0=10*rnorm(1), sigma.u0=10*runif(1),
 u1=10*rnorm(n.CA), mu.u1=10*rnorm(1), sigma.u1=10*runif(1),
 sigma.e=runif(1), .RNG.name='base::Wichmann-Hill', .RNG.seed=1987),
  "2"=list(u0=10*rnorm(n.CA), mu.u0=10*rnorm(1), sigma.u0=10*runif(1),
 u1=10*rnorm(n.CA), mu.u1=10*rnorm(1), sigma.u1=10*runif(1),
 sigma.e=runif(1), .RNG.name='base::Wichmann-Hill', .RNG.seed=2015),
  "3"=list(u0=10*rnorm(n.CA), mu.u0=10*rnorm(1), sigma.u0=10*runif(1),
 u1=10*rnorm(n.CA), mu.u1=10*rnorm(1), sigma.u1=10*runif(1),
 sigma.e=runif(1), .RNG.name='base::Wichmann-Hill', .RNG.seed=1234)))

# bundle the data
jags.dat5 <- list(y=usedat5$cases, x=usedat5$yr, CA=usedat5$CA,
```

```
    n.CA=n.CA, n.obs=dim(usedat5)[1], offs=log(usedat5$totalb))
str(jags.dat5)

set.seed(12345)

# initialise the model in JAGS
jags.m5 <- jags.model('m5.txt', jags.dat5, n.chains=3, inits=inits.m5,
 n.adapt=1000)
update(jags.m5,500)      #burn in
list.samplers(jags.m5)  #check which samplers are being used

# run length control (pilot run)
m5 <- coda.samples.dic(jags.m5, params5, n.iter=5000, thin=1)
raftery.diag(m5$samples)

# run coda.samples with DIC module
m5 <- coda.samples.dic(jags.m5, params5, n.iter=100000, thin=5)

# summarising the posterior distributions for parameters
summary(m5$samples)
m5$dic
effectiveSize(m5$samples)

# convergence diagnostics
gelman.diag(m5$samples)
geweke.diag(m5$samples)
heidel.diag(m5$samples)

## visually assessing convergence

# trace plots, mixing of chains
xyplot(m5$samples[][,1:5], strip=F, strip.left=T)
xyplot(m5$samples[][,6:11], strip=F, strip.left=T)

# shape of posterior distributions
densityplot(m5$samples[][,1:5], strip=F, strip.left=T)
densityplot(m5$samples[][,6:11], strip=F, strip.left=T)

# autocorrelation
acfplot(m5$samples[][,1:5], aspect='fill')
acfplot(m5$samples[][,6:11], aspect='fill')
```

A3. Choice of priors for estimation of variance parameters in hierarchical models

A half-Cauchy prior distribution can be coded in JAGS as described in Gelman and Pardoe[2]. For example, a half-Cauchy prior for the variance parameter sigma.u1 (i.e. the estimated SD of the random slopes across all CA subgroups) and a scale parameter S can be calculated as follows

- The half-Cauchy prior for `sigma.u1` is given by dividing a Normal distribution `z.u1` by the square root of a chi-squared distribution `chiSq.u1`

$$\text{sigma.u1} = \frac{\text{z.u1}}{\sqrt{\text{chiSq.u1}}}$$

- The precision of `z.u1` is

$$\text{zprec} = \left(\frac{1}{\text{S}}\right)^2$$

- The distribution is restricted to be greater than zero, since the SD cannot take negative values

```
z.u0 ~ Normal(0, zprec)I(0,)
```

- A chi-square distribution with 1 degree of freedom is used

```
chiSq.u0 ~ Gamma(0.5, 0.5)
```

This can similarly be coded in an rjags model using a truncated t distribution as follows

```
sigma.u1 ~ dt(0, zprec, 1)I(0,)
```

[2] Gelman A & Pardoe I (2006). Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics, 48, 241-251.

A4. Supplementary results for analysis of prevalence in neural tube defects

Table A5. Trends in neural tube defects from individual and frequentist hierarchical models (models 1-4).

Model	CAs modelled together	Random effect for registry	CA	Estimated trend (95% CI)	P-value	Over dispersion parameter	SD of registry random effect	SD of CA intercepts	SD of CA trends	cAIC
1	No	No	NTDs	-0.001 (-0.011, 0.010)	0.908	0				90.2
			Anencephaly	0.004 (-0.013, 0.021)	0.621	0				79.4
			Encephalocele	-0.021 (-0.055, 0.012)	0.212	0				66.5
			Spina Bifida	0.000 (-0.016, 0.015)	0.965	0				85.1
2	No	Yes	NTDs	-0.001 (-0.013, 0.010)	0.816	0.056	0.287			1063.2
			Anencephaly	0.004 (-0.014, 0.021)	0.691	0.079	0.422			885.7
			Encephalocele	-0.021 (-0.055, 0.012)	0.208	0	0.370			618.2
			Spina Bifida	-0.001 (-0.017, 0.015)	0.933	0.060	0.242			885.7
3	Yes	No	NTDs	-0.006 (-0.023, 0.012)	0.536	0	n/a	0.692 [a]	0.010 [a]	232.9
			Anencephaly	0.0002						
			Encephalocele	-0.019						
			Spina Bifida	0.003						
4	Yes	Yes	NTDs	-0.006 (-0.024, 0.011)	0.485	0.124	0.296	0.695 [a]	0.009 [a]	2466.9
			Anencephaly	-0.001						
			Encephalocele	-0.019						
			Spina Bifida	0.002						

[a] *Correlation between random intercepts and slopes estimated to be 1 in models 3 and 4*

Table A6. Notation in Bayesian hierarchical models for neural tube defects.

Parameter	Description
u0[1]	Random intercept for **Anencephaly** subgroup
u0[2]	Random intercept for **Encephalocele** subgroup
u0[3]	Random intercept for **Spina bifida** subgroup
mu.u0	Mean of random intercepts for NTD subgroups
sigma.u0	Standard deviation of random intercepts for NTD subgroups
u1[1]	Random slope (trend) for **Anencephaly** subgroup
u1[2]	Random slope (trend) for **Encephalocele** subgroup
u1[3]	Random slope (trend) for **Spina bifida** subgroup
mu.u1	Mean of random slopes for NTD subgroups
sigma.u1	Standard deviation of random slopes for NTD subgroups
r[j]	Random intercept for registry j
mu.r	Mean of random slopes for registries
sigma.r	Standard deviation of random slopes for registries
sigma.e	Standard deviation of overdispersion term

The Raftery-Lewis diagnostic

The Raftery-Lewis diagnostic for each model was based on a pilot run of 5000 samples with thin=1. The dependence factor estimates how much the autocorrelation inflates the required sample size, with mean values >5 (across all chains) indicating strong autocorrelation. The total estimated required sample size gives the approximate iterations required to estimate the 95% posterior confidence limits (i.e. the 0.025 and 0.975 quantiles, with probability=0.95) to have actual posterior probability within 0.0125 (the accuracy) of that estimated. Therefore, the 95% PCIs would have actual posterior probability between 0.925 and 0.975, with confidence limits from between the 0.0125–0.0375 quantile up to the 0.9625–0.9875 quantile.

Table A7. Raftery-Lewis diagnostic for Bayesian hierarchical model for neural tube defects pooled over registries (model 5).

Parameter	**Burn-in** (M)		**Total estimated required sample size** (N)		**Dependence factor**, $I=\frac{M+N}{N_{min}}$	
quantile	*0.025*	*0.975*	*0.025*	*0.975*	*0.025*	*0.975*
u1[1]	2.3	3.0	635	690	1.1	1.1
u1[2]	3.3	3.0	717	705	1.2	1.2
u1[3]	2.7	2.0	634	608	1.1	1.0
mu.u1	6.7	11.7	1360	2117	2.3	3.5
sigma.u1	3.3	11.7	721	1986	1.2	3.3
u0[1]	3.7	3.3	763	691	1.3	1.2
u0[2]	4.0	3.7	811	772	1.4	1.3
u0[3]	6.0	3.0	1157	701	1.9	1.2
mu.u0	19.0	18.3	4172	3869	7.0	6.5
sigma.u0	29.3	80.0	5331	14134	8.9	23.6
sigma.e	86.0	33.0	14124	5632	23.5	9.4

Table A8. Summary of posterior distribution for NTDs model 5.

Parameter	Mean	SD	2.5%	50%	97.5%	ESS [a]	PSRF [b]
u0[1]	-7.784	0.027	-7.837	-7.784	-7.731	60,545	1
u0[2]	-9.162	0.050	-9.261	-9.162	-9.065	58,569	1
u0[3]	-7.612	0.025	-7.662	-7.612	-7.563	60,046	1
mu.u0	-8.160	1.680	-11.690	-8.176	-4.499	60,844	1
sigma.u0	2.282	1.907	0.525	1.577	7.969	28,850	1
u1[1]	0.002	0.009	-0.014	0.002	0.020	40,700	1
u1[2]	-0.012	0.015	-0.046	-0.010	0.013	22,144	1
u1[3]	-0.001	0.008	-0.017	-0.001	0.015	45,818	1
mu.u1	-0.004	0.139	-0.118	-0.002	0.102	80,276	1.18
sigma.u1	0.070	0.234	0.001	0.018	0.508	471	1.2
sigma.e	0.029	0.021	0.001	0.025	0.079	2,407	1

[a] *Effective sample size*
[b] *Estimated potential scale reduction factor*

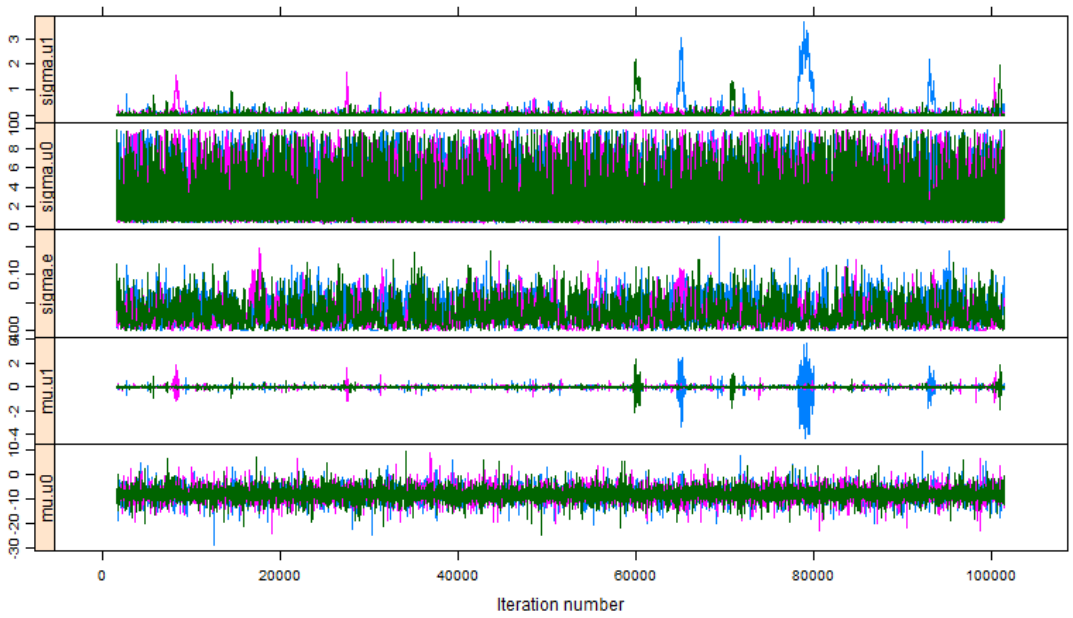


Figure A1. Trace plots for NTDs model 5: group level parameters.

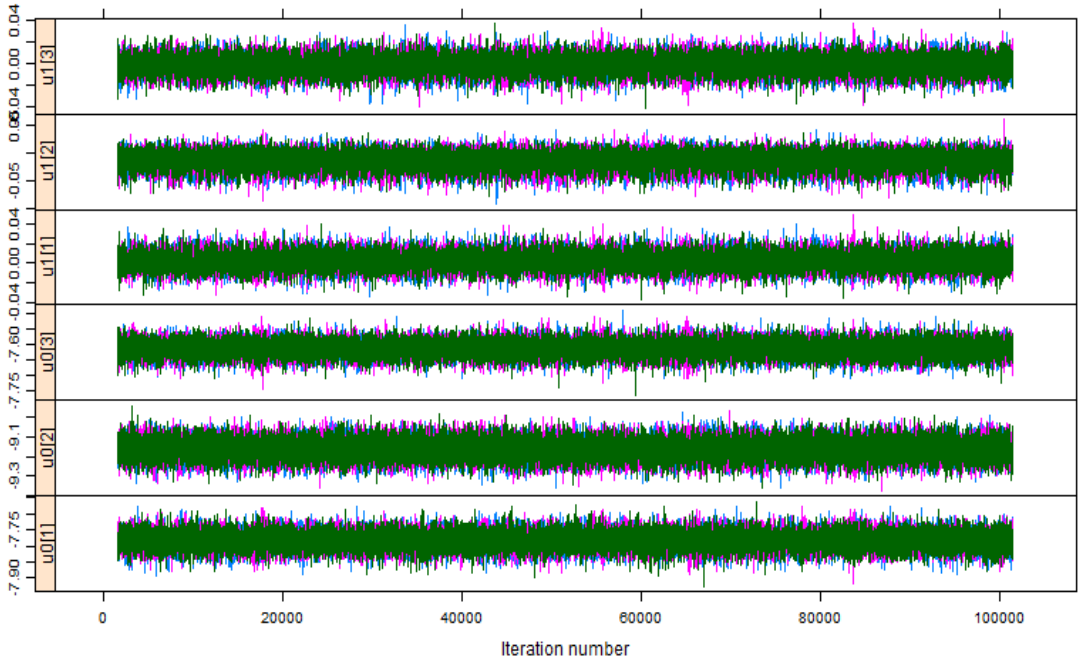


Figure A2. Trace plots for model 5: random intercepts and slopes for CA subgroups

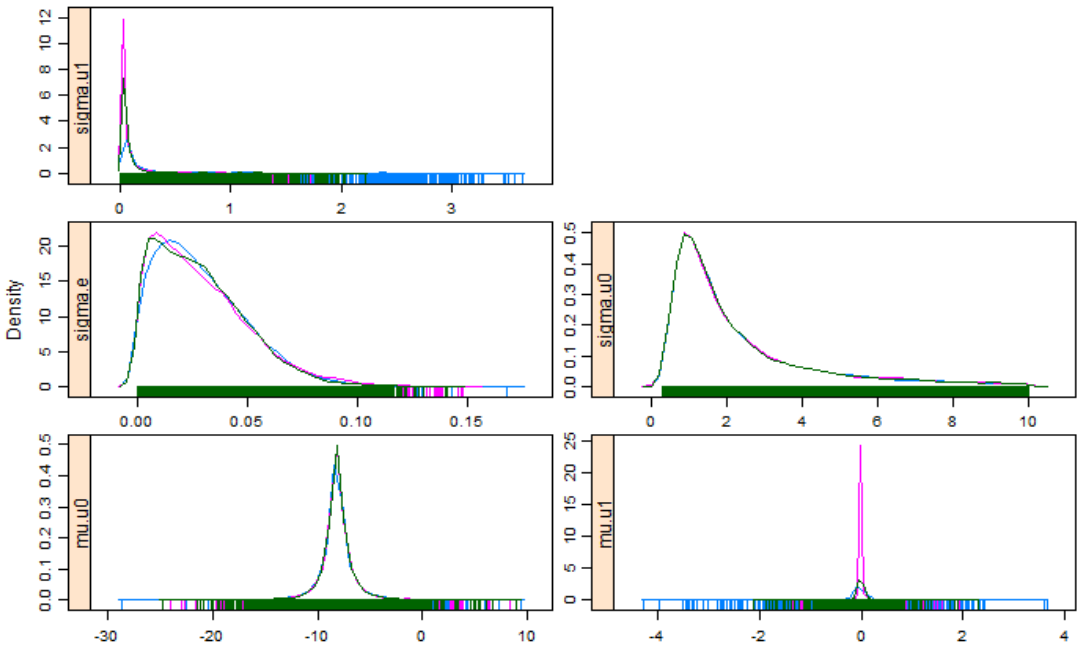


Figure A3. Density plots for NTDs model 5: group level parameters.

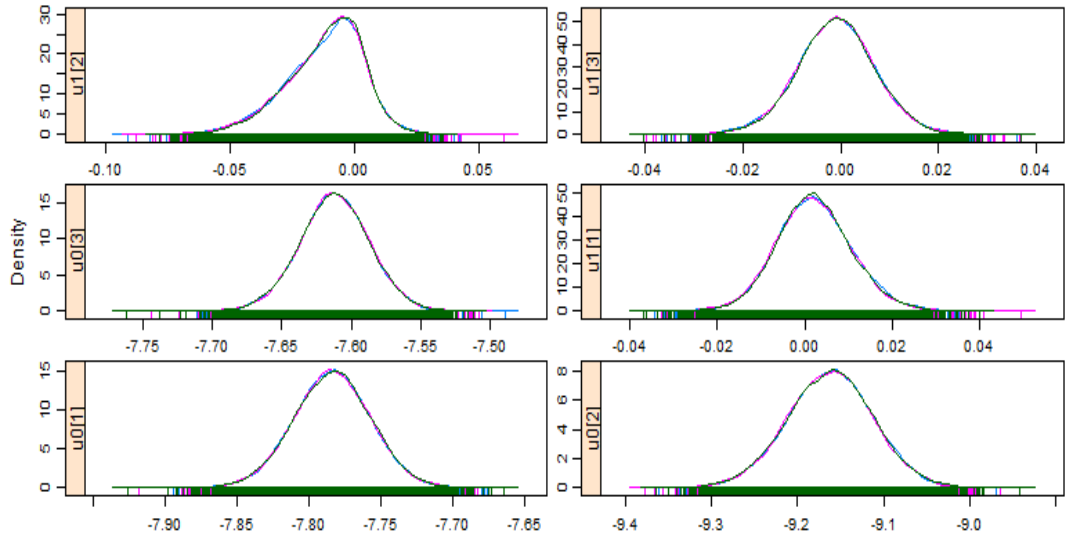


Figure A4. Density plots for model 5: random intercepts and slopes for CA subgroups.

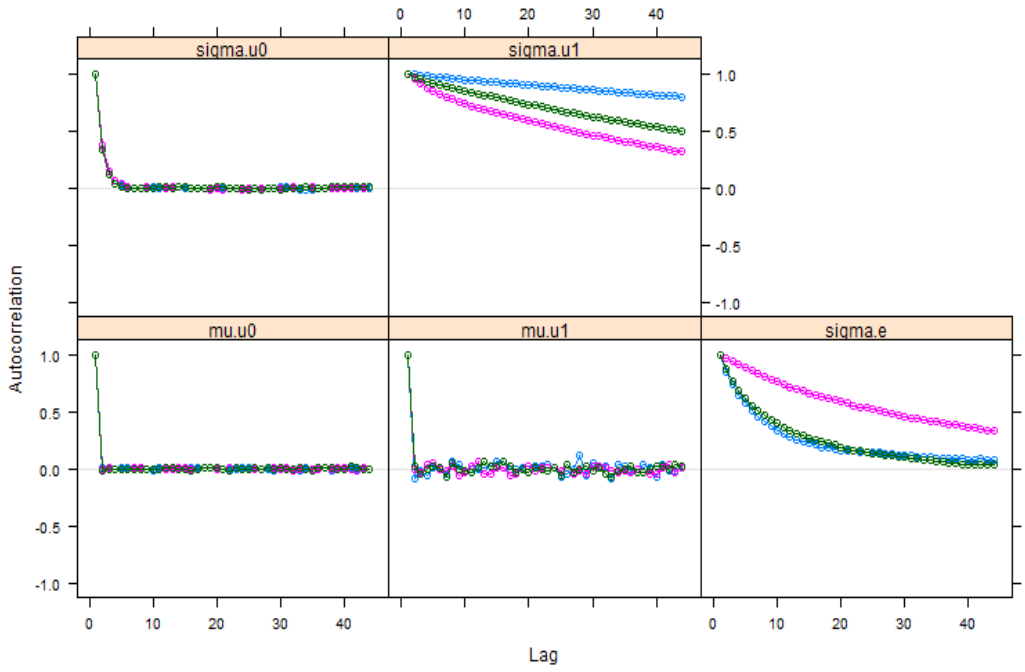


Figure A5. Autocorrelation function plots for NTDs model 5: group level parameters.

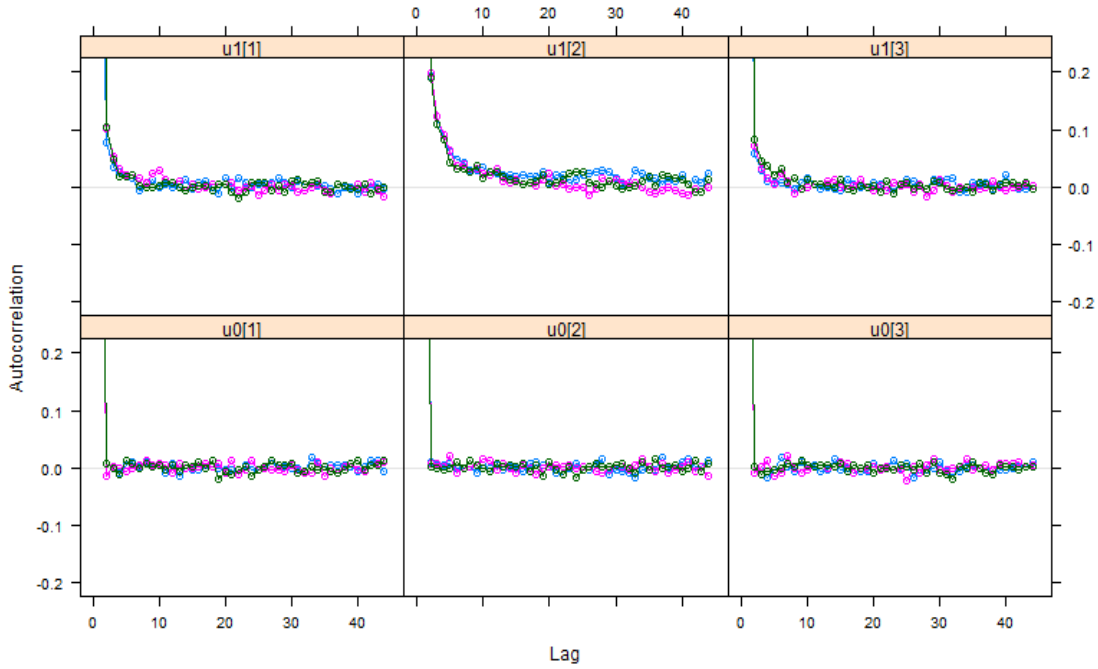


Figure A6. Autocorrelation function plots for NTDs model 5: random intercepts and slopes for CA subgroups.

Table A9. Raftery-Lewis diagnostic for model 6A.

Parameter	**Burn-in** (M)		**Total estimated required sample size** (N)		**Dependence factor**, $I=\frac{M+N}{N_{min}}$	
quantile	*0.025*	*0.975*	*0.025*	*0.975*	*0.025*	*0.975*
u1[1]	8.3	7.3	1444	1424	2.4	2.4
u1[2]	18.7	12.0	3293	2200	5.5	3.7
u1[3]	6.3	8.7	1142	1676	1.9	2.8
mu.u1	20.0	20.7	3310	4903	5.5	8.2
sigma.u1	36.0	75.3	6279	13094	10.4	21.8
u0[1]	105.0	78.0	17772	13043	29.6	21.7
u0[2]	119.3	70.0	20664	12091	34.5	20.2
u0[3]	102.3	89.7	16996	14622	28.3	24.4
mu.u0	7.0	9.0	1345	1775	2.2	3.0
sigma.u0	3.3	13.7	750	2293	1.3	3.8
r[1]	77.7	111.0	13418	19120	22.4	31.9
r[2]	87.7	111.3	15043	19676	25.1	32.8
r[3]	78.0	96.7	13065	16319	21.8	27.2
r[4]	77.7	70.3	13331	12025	22.2	20.0
r[5]	110.3	102.3	18965	17867	31.6	29.8
r[6]	76.3	85.0	13380	14633	22.3	24.4
r[7]	100.7	160.0	16807	29071	28.0	48.6
r[8]	91.3	96.3	15448	15918	25.7	26.5
r[9]	75.0	170.7	13018	29811	21.7	49.7
r[10]	80.0	89.0	13962	15173	23.3	25.3
r[11]	107.3	119.7	18140	20614	30.2	34.3
r[12]	84.3	93.0	14535	16229	24.2	27.1
r[13]	144.0	131.3	24926	21930	41.5	36.6
r[14]	83.3	122.7	13739	20757	22.9	34.6
r[15]	81.7	166.0	13691	30404	22.8	50.8
r[16]	87.3	123.7	15383	21040	25.6	35.1
r[17]	111.7	119.0	17977	19595	29.9	32.7
r[18]	91.7	112.3	15447	19397	25.8	32.3
mu.r	60.0	100.7	10447	17532	17.4	29.2
sigma.r	4.7	7.7	862	1326	1.4	2.2
sigma.e	62.0	50.3	10450	8322	17.4	13.9

Table A10. Summary of posterior distribution for model 6A.

Parameter	Mean	SD	2.5%	50%	97.5%	ESS	PSRF [a]
u0[1]	-4.868	19.379	-38.700	-2.324	27.959	11.6	4.93
u0[2]	-6.236	19.380	-40.070	-3.687	26.587	11.7	4.93
u0[3]	-4.684	19.379	-38.530	-2.133	28.143	11.6	4.93
mu.u0	-5.253	19.401	-39.170	-2.763	27.687	12.6	4.83
sigma.u0	2.311	1.935	0.523	1.598	8.038	27661.0	1
u1[1]	0.002	0.009	-0.015	0.001	0.019	30183.4	1
u1[2]	-0.012	0.015	-0.047	-0.010	0.013	16141.1	1
u1[3]	-0.002	0.008	-0.018	-0.002	0.013	32577.6	1
mu.u1	-0.005	0.102	-0.105	-0.003	0.089	101156.7	1.16
sigma.u1	0.054	0.145	0.001	0.018	0.374	2907.4	1.13
r[1]	-3.141	19.379	-35.980	-5.700	30.698	11.6	4.93
r[2]	-2.895	19.378	-35.720	-5.442	30.927	11.8	4.93
r[3]	-2.883	19.379	-35.720	-5.427	30.945	11.7	4.93
r[4]	-3.489	19.379	-36.300	-6.034	30.364	11.6	4.93
r[5]	-3.462	19.379	-36.290	-6.010	30.390	11.6	4.93
r[6]	-3.256	19.378	-36.070	-5.798	30.573	11.7	4.93
r[7]	-3.041	19.379	-35.870	-5.591	30.805	11.7	4.93
r[8]	-3.220	19.379	-36.070	-5.766	30.599	11.5	4.93
r[9]	-2.967	19.378	-35.790	-5.513	30.875	11.5	4.93
r[10]	-3.076	19.379	-35.890	-5.609	30.739	11.5	4.93
r[11]	-3.344	19.379	-36.170	-5.888	30.497	11.6	4.93
r[12]	-2.838	19.379	-35.660	-5.386	30.988	11.6	4.93
r[13]	-2.679	19.379	-35.510	-5.229	31.161	11.6	4.93
r[14]	-2.411	19.379	-35.250	-4.952	31.435	11.6	4.93
r[15]	-2.916	19.379	-35.740	-5.468	30.906	11.6	4.93
r[16]	-2.828	19.379	-35.660	-5.371	31.012	11.5	4.93
r[17]	-2.844	19.379	-35.670	-5.392	30.986	11.6	4.93
r[18]	-2.662	19.378	-35.490	-5.207	31.185	11.5	4.93
mu.r	-2.997	19.378	-35.840	-5.543	30.842	11.5	4.93
sigma.r	0.322	0.066	0.221	0.313	0.477	49262.5	1
sigma.e	0.120	0.038	0.040	0.122	0.188	965.8	1.02

[a] *Estimated potential scale reduction factor*

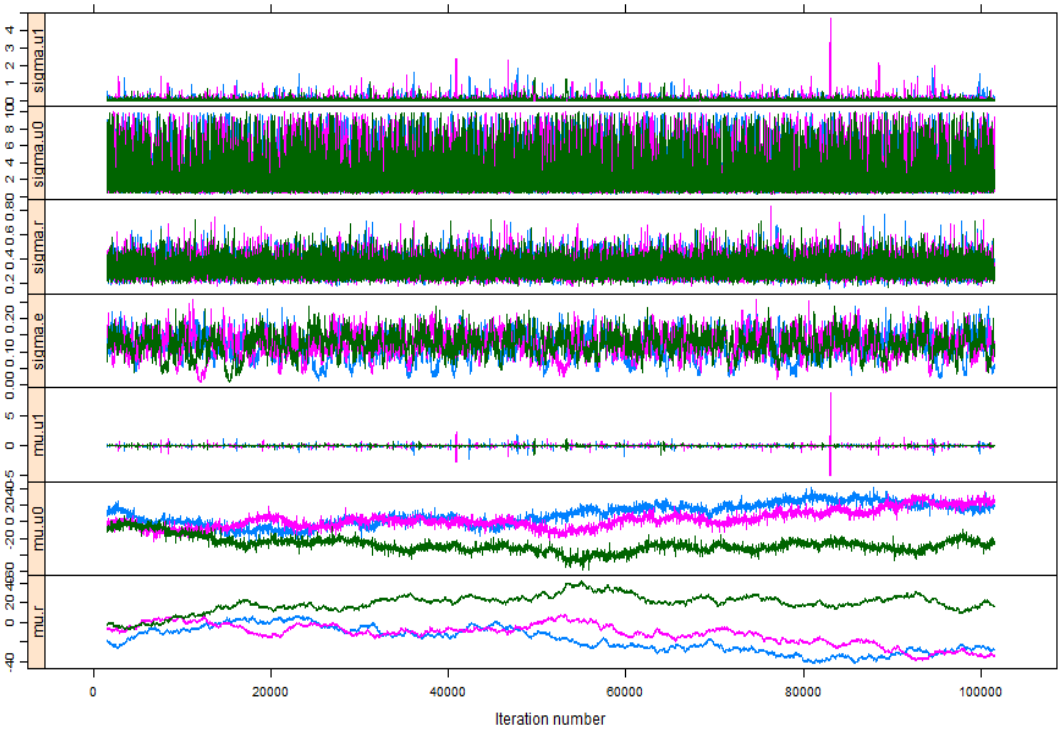


Figure A7. *Trace plots for model 6a: group level parameters.*

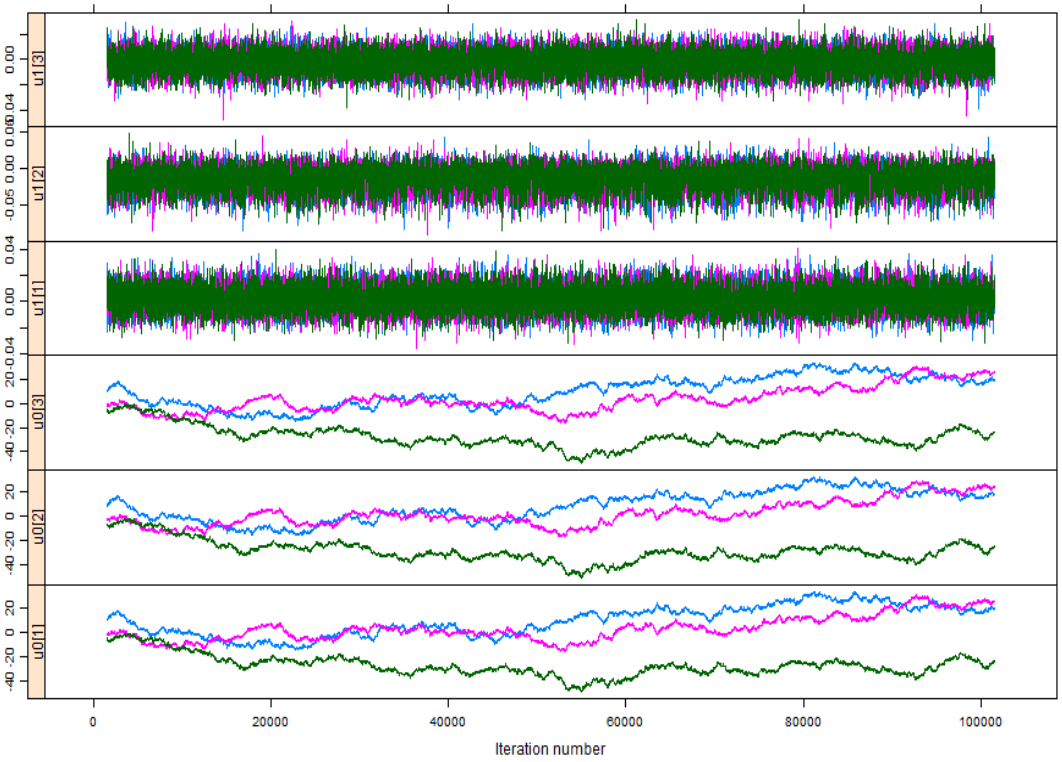


Figure A8. Trace plots for model 6a: random intercepts and slopes for CAs.

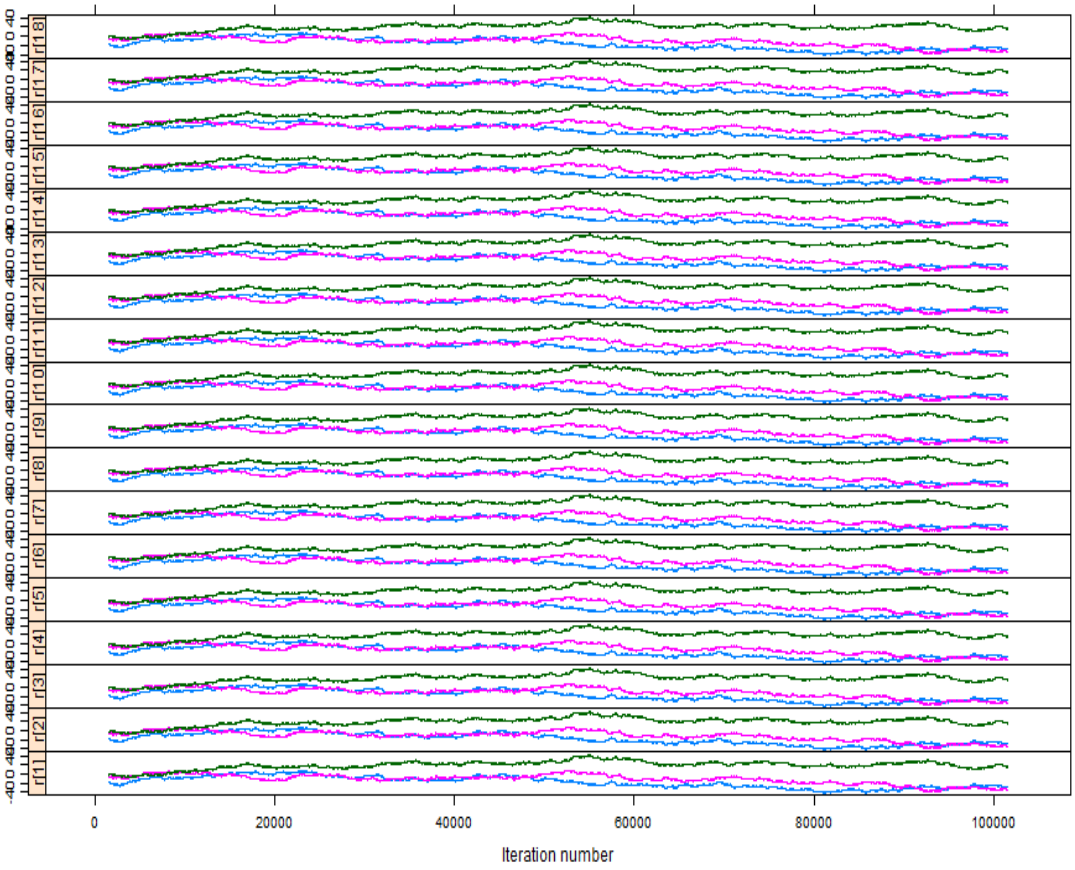


Figure A9. Trace plots for model 6A: random intercepts for registry.

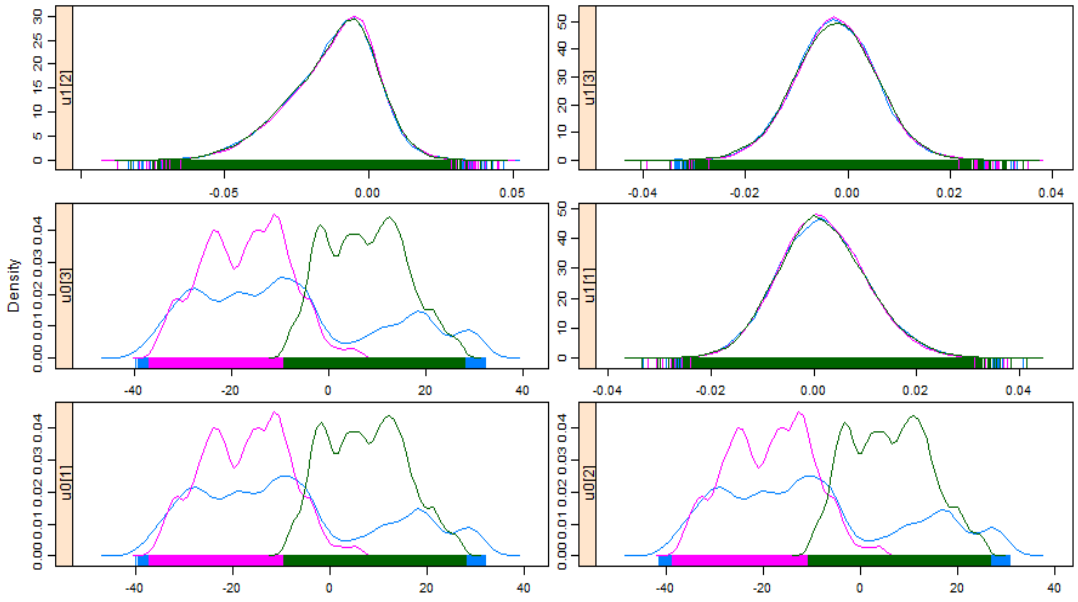


Figure A10. Density plots for model 6A: group level parameters.

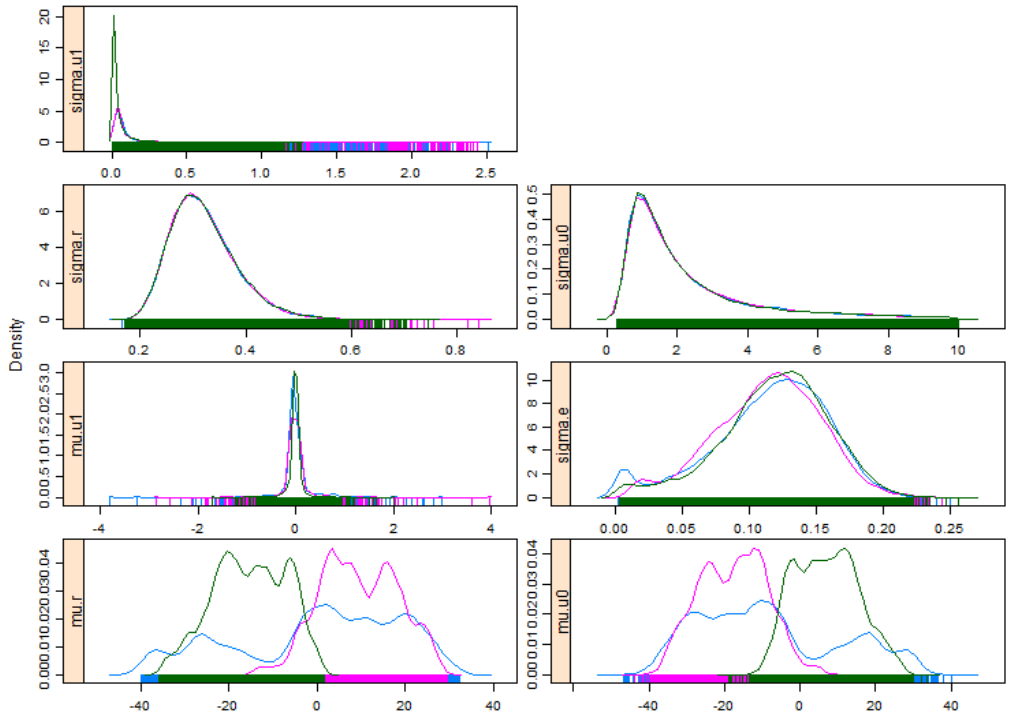


Figure A11. Density plots for model 6A: random intercepts and slopes for CAs.

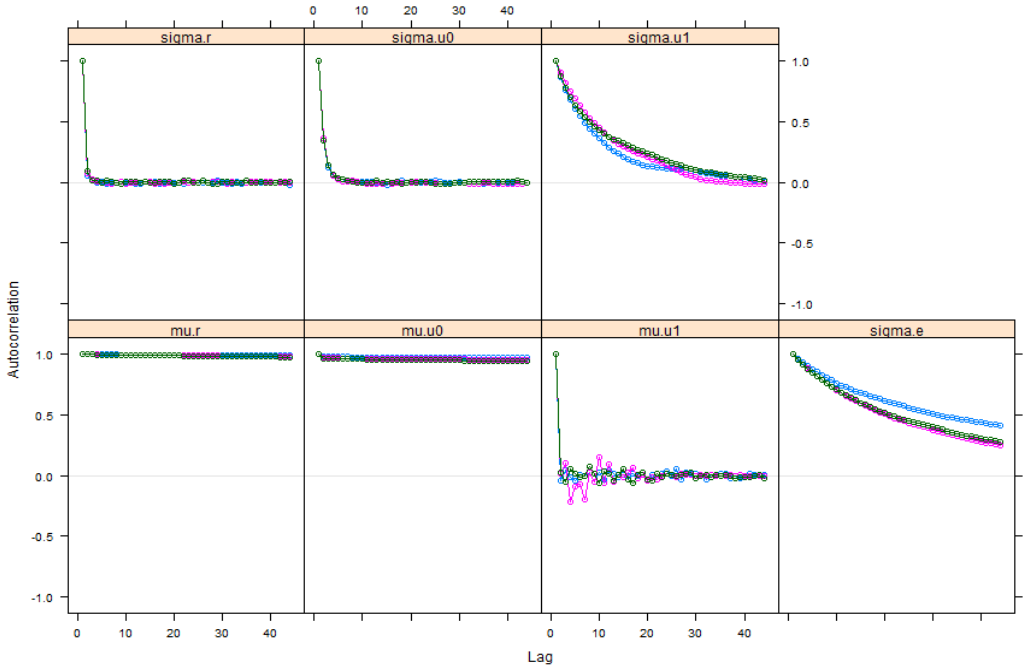


Figure A12. Autocorrelation function plots for model 6A: group level parameters.

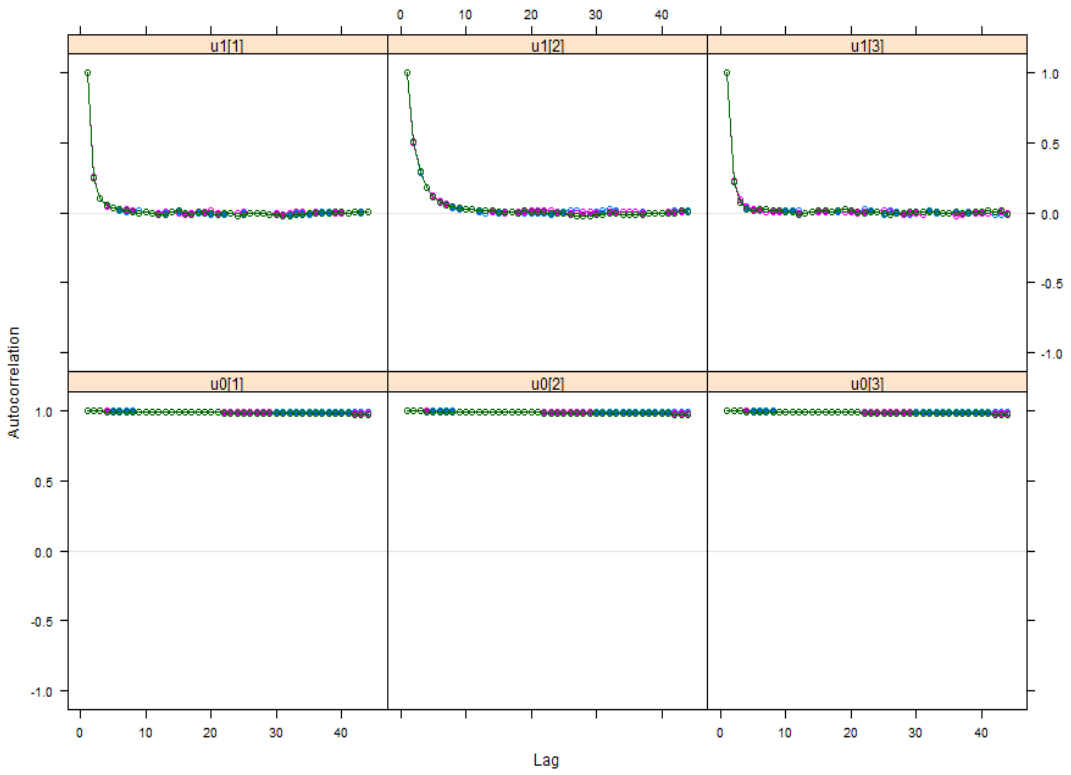


Figure A13. Autocorrelation function plots for model 6A: random intercepts and slopes for CAs.

Table A11. Raftery-Lewis diagnostic for model 6B.

Parameter	**Burn-in** (M)		**Total estimated required sample size** (N)		**Dependence factor**, $I=\frac{M+N}{N_{min}}$	
quantile	*0.025*	*0.975*	*0.025*	*0.975*	*0.025*	*0.975*
u1[1]	7.3	6.7	1293	1207	2.2	2.0
u1[2]	18.0	11.3	3239	2053	5.4	3.4
u1[3]	7.7	7.3	1554	1422	2.6	2.4
mu.u1	10.0	10.0	1940	1949	3.2	3.2
sigma.u1	34.0	75.3	5825	13060	9.7	21.8
u0[1], r[1]	7.7	10.0	1421	2013	2.4	3.4
u0[1], r[2]	15.0	13.3	2611	2562	4.4	4.3
u0[1], r[3]	7.7	8.0	1356	1645	2.3	2.7
u0[1], r[4]	6.7	9.0	1345	1675	2.2	2.8
u0[1], r[5]	13.3	12.0	2308	2147	3.9	3.6
u0[1], r[6]	6.7	9.0	1137	1675	1.9	2.8
u0[1], r[7]	13.3	11.3	2549	2146	4.2	3.6
u0[1], r[8]	16.7	14.0	2889	2375	4.8	4.0
u0[1], r[9]	8.3	10.7	1519	1906	2.5	3.2
u0[1], r[10]	12.0	7.3	2137	1340	3.6	2.2
u0[1], r[11]	18.0	10.0	2967	1837	4.9	3.1
u0[1], r[12]	7.3	8.7	1513	1670	2.5	2.8
u0[1], r[13]	4.0	3.7	784	746	1.3	1.2
u0[1], r[14]	7.0	8.3	1220	1567	2.0	2.6
u0[1], r[15]	3.3	3.3	737	713	1.2	1.2
u0[1], r[16]	12.7	8.7	2335	1604	3.9	2.7
u0[1], r[17]	15.7	16.3	2671	2965	4.5	4.9
u0[1], r[18]	7.3	8.7	1300	1625	2.2	2.7
mu.u0[1]	3.0	2.0	664	611	1.1	1.0
sigma.u0[1]	5.7	7.0	1116	1281	1.9	2.1
u0[2], r[1]	5.3	7.0	1033	1422	1.7	2.4
u0[2], r[2]	9.3	11.7	1840	2150	3.1	3.6
u0[2], r[3]	4.3	4.3	991	863	1.7	1.4
u0[2], r[4]	12.0	9.3	2028	1670	3.4	2.8
u0[2], r[5]	27.7	14.7	5260	2653	8.8	4.4
u0[2], r[6]	7.3	8.3	1530	1454	2.6	2.4
u0[2], r[7]	13.3	15.7	2346	2491	3.9	4.2
u0[2], r[8]	19.3	20.0	3494	3652	5.8	6.1
u0[2], r[9]	10.3	11.0	1822	2078	3.0	3.5
u0[2], r[10]	7.3	5.7	1317	1104	2.2	1.8
u0[2], r[11]	10.0	11.3	1766	1982	2.9	3.3
u0[2], r[12]	6.0	7.0	1048	1325	1.7	2.2
u0[2], r[13]	8.0	11.3	1500	2046	2.5	3.4
u0[2], r[14]	10.7	15.3	1814	2853	3.0	4.8
u0[2], r[15]	6.7	6.3	1198	1084	2.0	1.8
u0[2], r[16]	11.3	8.0	2036	1528	3.4	2.6

u0[2], r[17]	10.0	12.7	1660	2380	2.8	4.0
u0[2], r[18]	6.0	7.0	1111	1315	1.9	2.2
mu.u0[2]	4.7	3.0	938	686	1.6	1.1
sigma.u0[2]	15.7	10.7	2690	1885	4.5	3.1
u0[3], r[1]	13.7	12.7	2300	2437	3.8	4.1
u0[3], r[2]	14.0	15.3	2767	2674	4.6	4.5
u0[3], r[3]	10.3	9.0	1863	1623	3.1	2.7
u0[3], r[4]	7.0	7.3	1287	1281	2.1	2.1
u0[3], r[5]	14.0	15.0	2660	2799	4.4	4.7
u0[3], r[6]	6.3	7.7	1124	1380	1.9	2.3
u0[3], r[7]	11.3	9.0	2224	1665	3.7	2.8
u0[3], r[8]	16.0	11.7	2925	2020	4.9	3.4
u0[3], r[9]	7.7	6.7	1406	1247	2.3	2.1
u0[3], r[10]	14.7	12.0	2589	2237	4.3	3.7
u0[3], r[11]	20.3	16.7	3642	2893	6.1	4.8
u0[3], r[12]	11.7	13.3	2056	2406	3.4	4.0
u0[3], r[13]	6.0	5.0	1079	908	1.8	1.5
u0[3], r[14]	10.7	10.7	1953	1843	3.3	3.1
u0[3], r[15]	5.3	4.3	944	980	1.6	1.6
u0[3], r[16]	9.3	8.0	1827	1556	3.0	2.6
u0[3], r[17]	14.0	12.0	2288	2133	3.8	3.6
u0[3], r[18]	8.0	6.3	1485	1140	2.5	1.9
mu.u0[3]	2.3	2.7	639	654	1.1	1.1
sigma.u0[3]	7.0	10.0	1458	1800	2.4	3.0
mu.ar	7.3	7.7	1483	1698	2.5	2.8
sigma.ar	3.7	12.7	741	2324	1.2	3.9
sigma.e	108.3	94.0	20442	16028	34.1	26.7

Table A12. Summary of posterior distribution for model 6B.

Parameter	Mean	SD	2.5%	50%	97.5%	ESS	PSRF [a]
u1[1]	0.002	0.008	-0.014	0.001	0.018	31311.4	1
u1[2]	-0.012	0.015	-0.046	-0.010	0.013	15622.4	1
u1[3]	-0.001	0.007	-0.016	-0.001	0.013	32375.1	1
mu.u1	-0.004	0.110	-0.102	-0.003	0.093	97477.3	1.08
sigma.u1	0.059	0.176	0.001	0.018	0.391	595.6	1.11
u0[1], r[1]	-8.258	0.136	-8.532	-8.255	-8.000	32261.8	1
u0[1], r[2]	-9.629	0.242	-10.140	-9.615	-9.188	19185.4	1
u0[1], r[3]	-7.717	0.100	-7.920	-7.715	-7.525	38498.4	1
u0[1], r[4]	-7.608	0.105	-7.818	-7.607	-7.409	38077.9	1
u0[1], r[5]	-9.524	0.235	-10.010	-9.512	-9.091	20311.4	1
u0[1], r[6]	-7.779	0.106	-7.993	-7.778	-7.577	40597.2	1
u0[1], r[7]	-7.671	0.143	-7.958	-7.669	-7.397	32076.4	1
u0[1], r[8]	-9.148	0.250	-9.658	-9.141	-8.674	20885.0	1
u0[1], r[9]	-7.566	0.125	-7.818	-7.565	-7.325	36731.8	1
u0[1], r[10]	-8.383	0.129	-8.642	-8.381	-8.137	33251.8	1
u0[1], r[11]	-9.531	0.209	-9.962	-9.523	-9.146	22416.4	1
u0[1], r[12]	-8.073	0.109	-8.292	-8.070	-7.865	36793.2	1
u0[1], r[13]	-7.636	0.059	-7.752	-7.635	-7.522	57353.9	1
u0[1], r[14]	-9.227	0.116	-9.459	-9.225	-9.005	35038.9	1
u0[1], r[15]	-7.553	0.056	-7.664	-7.552	-7.445	59382.5	1
u0[1], r[16]	-8.060	0.151	-8.367	-8.056	-7.776	29228.7	1
u0[1], r[17]	-9.266	0.239	-9.757	-9.258	-8.818	19427.2	1
u0[1], r[18]	-7.789	0.125	-8.041	-7.786	-7.551	35170.6	1
mu.u0[1]	-7.932	0.119	-8.173	-7.931	-7.700	52345.6	1
sigma.u0[1]	0.475	0.100	0.320	0.461	0.709	37986.0	1
u0[2], r[1]	-7.470	0.076	-7.621	-7.469	-7.324	48266.3	1
u0[2], r[2]	-8.989	0.147	-9.286	-8.987	-8.706	27880.6	1
u0[2], r[3]	-7.392	0.072	-7.536	-7.391	-7.253	49499.3	1
u0[2], r[4]	-8.235	0.140	-8.520	-8.232	-7.970	31909.5	1
u0[2], r[5]	-9.987	0.320	-10.680	-9.963	-9.433	12221.4	1
u0[2], r[6]	-7.719	0.106	-7.930	-7.717	-7.518	37688.9	1
u0[2], r[7]	-7.851	0.201	-8.262	-7.846	-7.471	21073.2	1
u0[2], r[8]	-9.169	0.302	-9.790	-9.161	-8.596	15916.1	1
u0[2], r[9]	-7.554	0.157	-7.869	-7.552	-7.252	30816.0	1
u0[2], r[10]	-7.713	0.092	-7.898	-7.712	-7.537	40528.9	1
u0[2], r[11]	-9.047	0.165	-9.380	-9.044	-8.733	26608.8	1
u0[2], r[12]	-7.627	0.086	-7.798	-7.627	-7.462	41043.0	1
u0[2], r[13]	-7.329	0.103	-7.536	-7.327	-7.132	38087.4	1
u0[2], r[14]	-8.710	0.190	-9.090	-8.707	-8.347	20233.8	1
u0[2], r[15]	-7.105	0.092	-7.290	-7.105	-6.928	37884.8	1
u0[2], r[16]	-8.495	0.163	-8.826	-8.490	-8.189	29355.5	1
u0[2], r[17]	-8.969	0.192	-9.353	-8.966	-8.603	23426.7	1
u0[2], r[18]	-7.541	0.100	-7.740	-7.540	-7.347	41381.6	1

mu.u0[2]	-9.212	0.119	-9.454	-9.209	-8.980	34869.8	1
sigma.u0[2]	0.425	0.114	0.243	0.410	0.688	16673.7	1
u0[3], r[1]	-8.745	0.233	-9.227	-8.736	-8.315	20256.5	1
u0[3], r[2]	-8.974	0.231	-9.440	-8.970	-8.531	20435.5	1
u0[3], r[3]	-7.848	0.139	-8.131	-7.844	-7.586	31753.2	1
u0[3], r[4]	-7.686	0.094	-7.875	-7.685	-7.503	40918.5	1
u0[3], r[5]	-9.264	0.184	-9.642	-9.260	-8.917	23751.8	1
u0[3], r[6]	-7.656	0.089	-7.832	-7.655	-7.486	42033.4	1
u0[3], r[7]	-8.511	0.127	-8.766	-8.509	-8.270	36034.4	1
u0[3], r[8]	-9.572	0.200	-9.989	-9.565	-9.202	21778.1	1
u0[3], r[9]	-8.034	0.100	-8.235	-8.032	-7.844	37998.1	1
u0[3], r[10]	-7.993	0.181	-8.361	-7.989	-7.652	23423.4	1
u0[3], r[11]	-8.942	0.254	-9.444	-8.939	-8.450	19157.7	1
u0[3], r[12]	-7.769	0.149	-8.073	-7.765	-7.485	28144.5	1
u0[3], r[13]	-7.620	0.080	-7.778	-7.619	-7.466	42596.3	1
u0[3], r[14]	-8.689	0.130	-8.951	-8.687	-8.442	31516.3	1
u0[3], r[15]	-7.393	0.071	-7.533	-7.392	-7.257	50857.1	1
u0[3], r[16]	-7.468	0.080	-7.630	-7.467	-7.314	42300.5	1
u0[3], r[17]	-9.314	0.177	-9.673	-9.310	-8.978	25512.5	1
u0[3], r[18]	-7.680	0.085	-7.850	-7.679	-7.518	41675.0	1
mu.u0[3]	-7.658	0.071	-7.799	-7.657	-7.518	51122.2	1
sigma.u0[3]	0.272	0.060	0.178	0.264	0.411	41844.3	1
mu.ar**	-8.241	1.719	-11.820	-8.253	-4.503	61471.2	1
sigma.ar**	2.268	1.916	0.504	1.559	8.014	30373.1	1
sigma.e	0.064	0.039	0.002	0.063	0.140	239.6	1.02

**Estimated potential scale reduction factor*

***mu.ar is the mean intercept across all CA*registry combinations, and sigma.ar the SD of the intercepts across all CA*registry combinations*

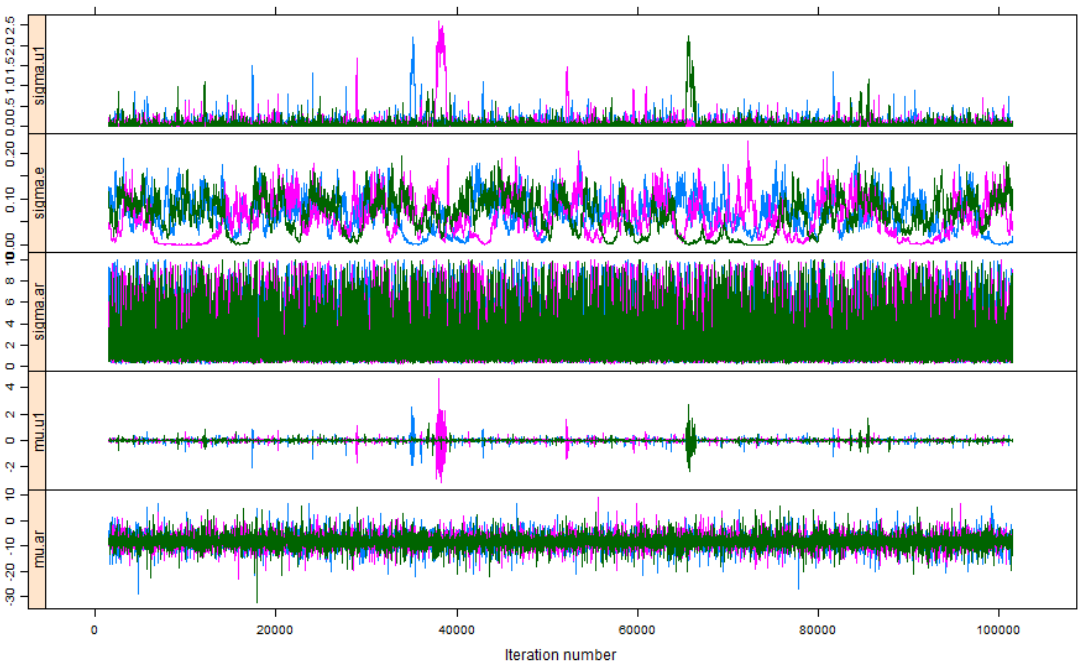


Figure A14. Trace plots for model 6b: group level parameters.

Note that in Figure A14 mu.ar is the mean intercept across all CA-registry combinations, sigma.ar the SD of the intercepts across all CA-registry combinations, mu.u1 the mean slope and sigma.u1 the SD of the trends across the three NTD subgroups.

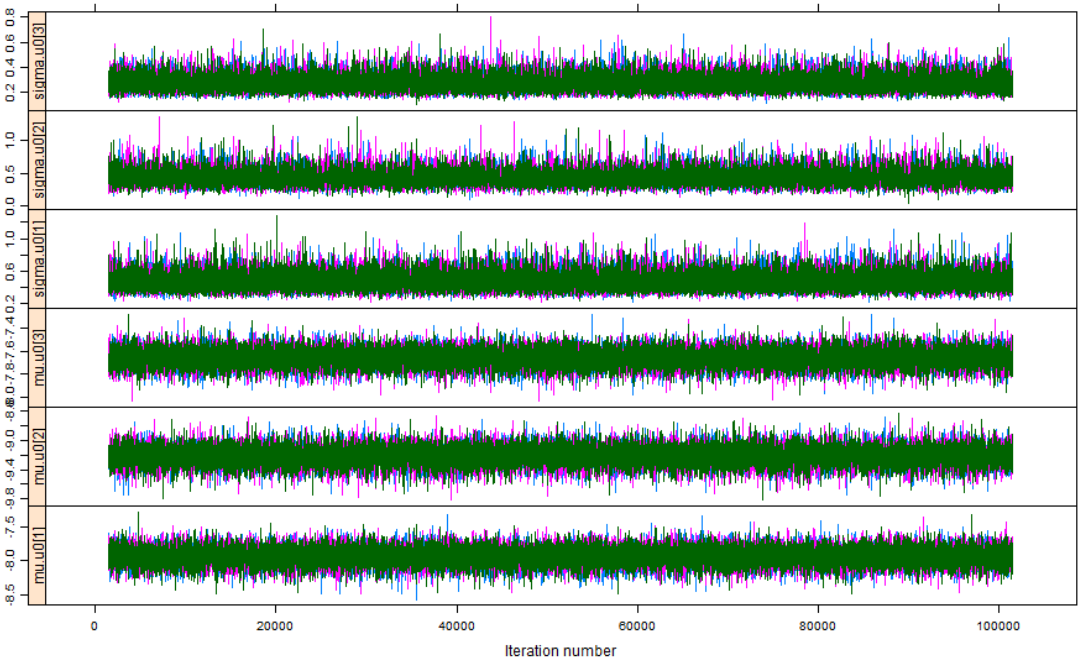


Figure A15. Trace plots for model 6B: mean and SD for random intercepts in each CA subgroup.

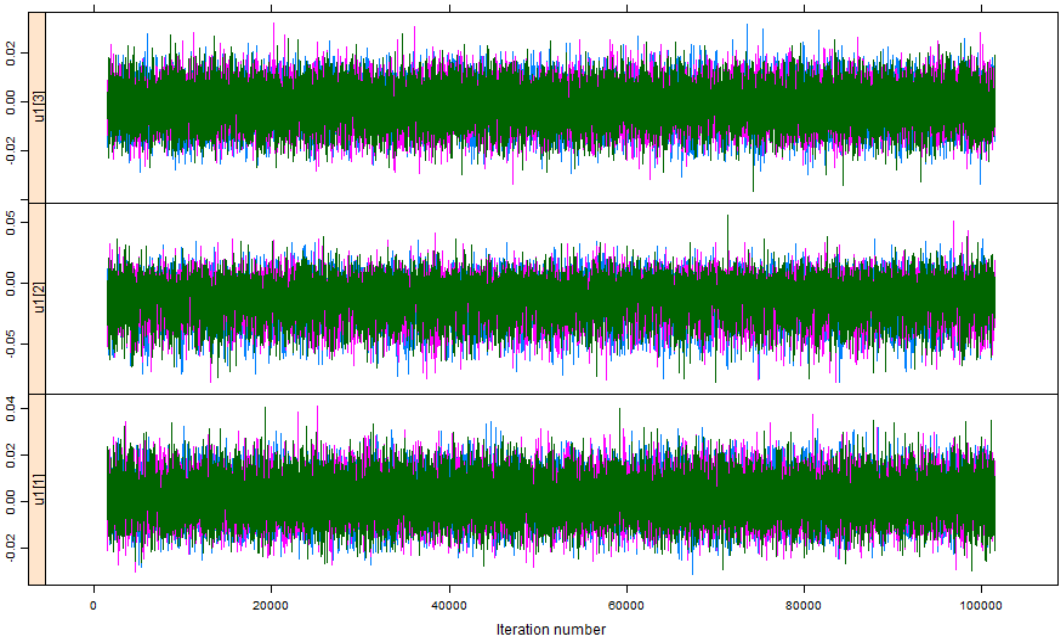


Figure A16. Trace plots for model 6B: random slope in each CA subgroup.

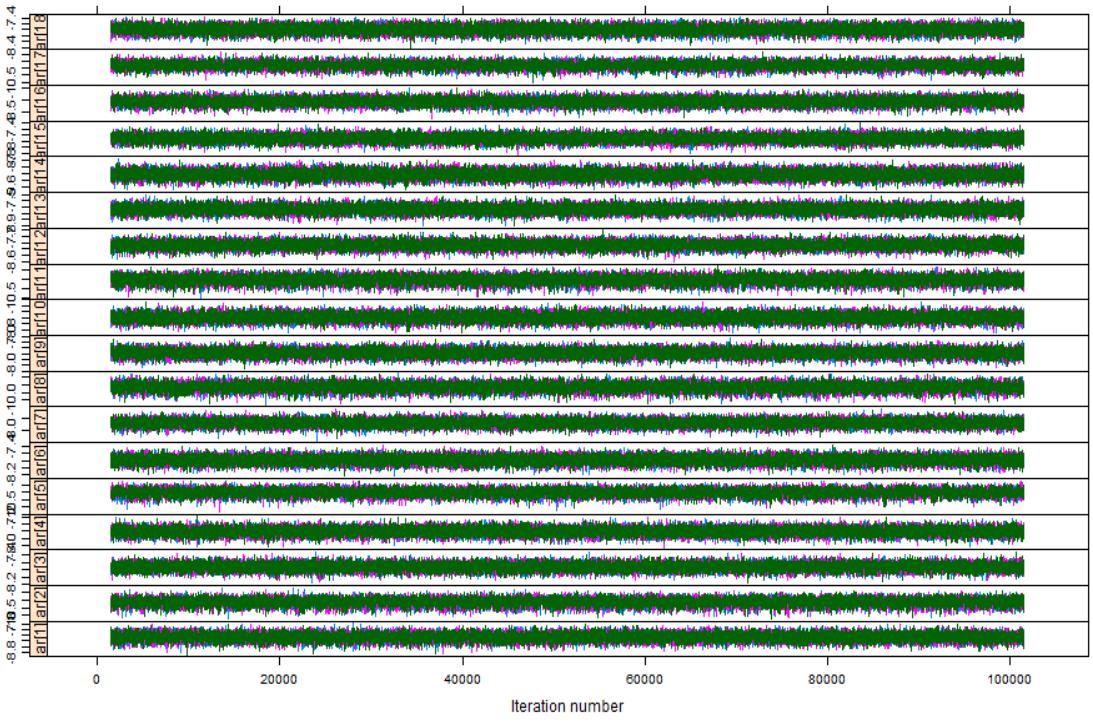


Figure A17. Trace plots for model 6b: random intercepts for intercepts for registries in the first CA subgroup (Note that the trace plots look the same for the registries for the other two CA subgroups).

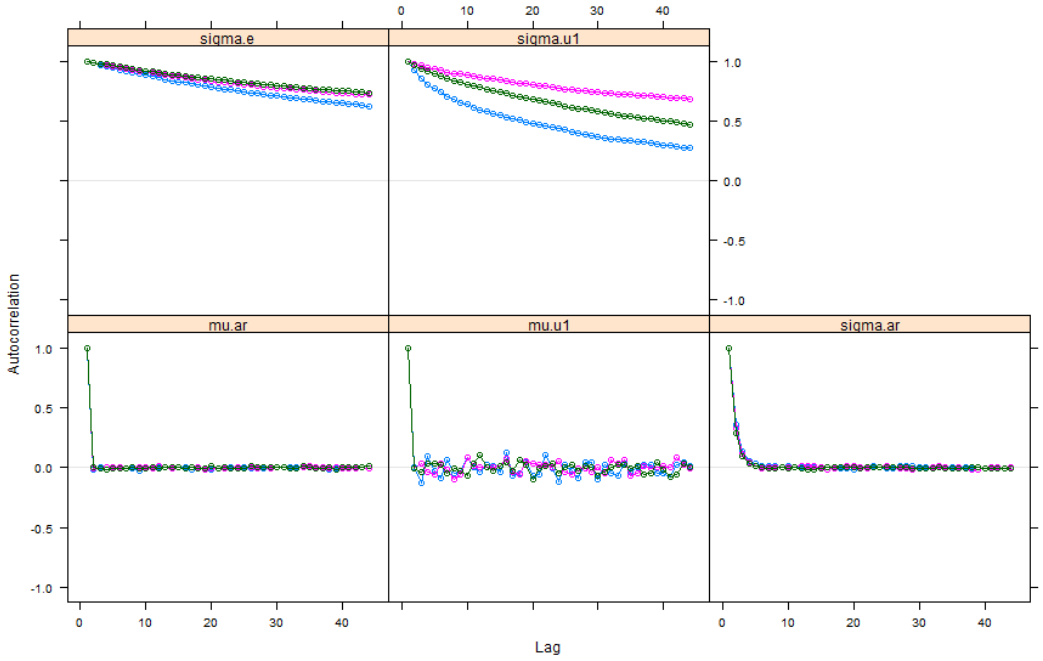


Figure A18. Autocorrelation function plots for model 6b: group level parameters.

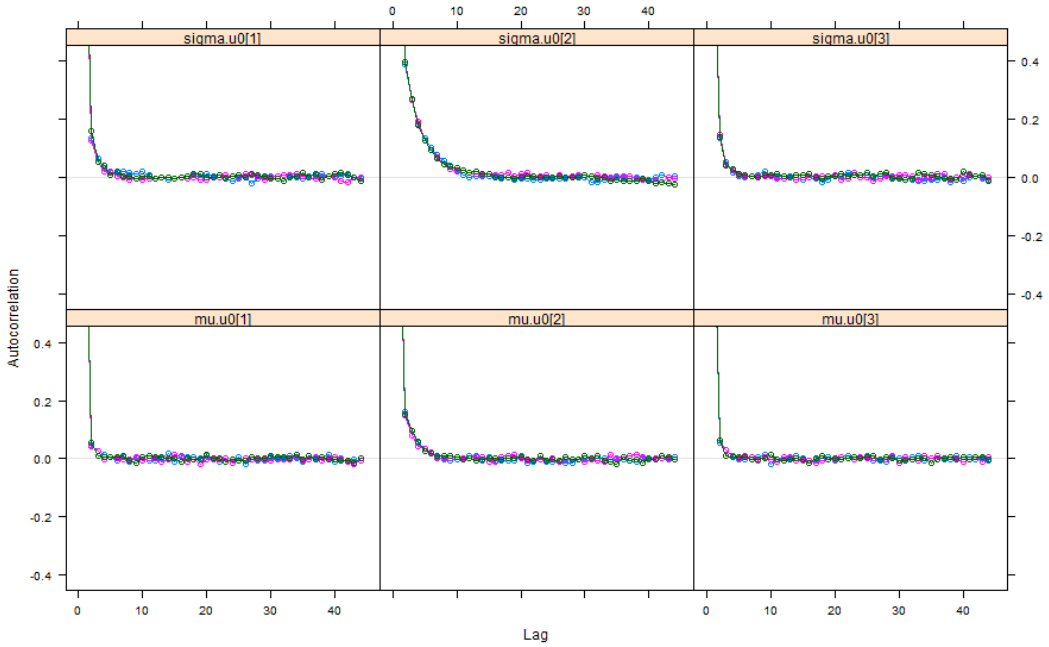


Figure A19. Autocorrelation function plots for model 6b: mean and SD for random intercepts in each CA subgroup.

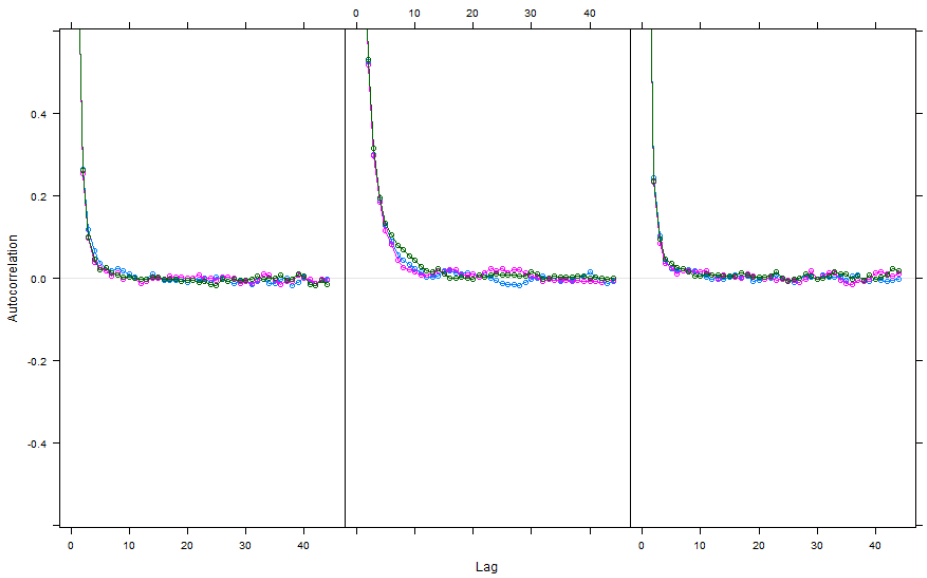


Figure A20. Autocorrelation function plots for model 6b: random slope in each CA subgroup.

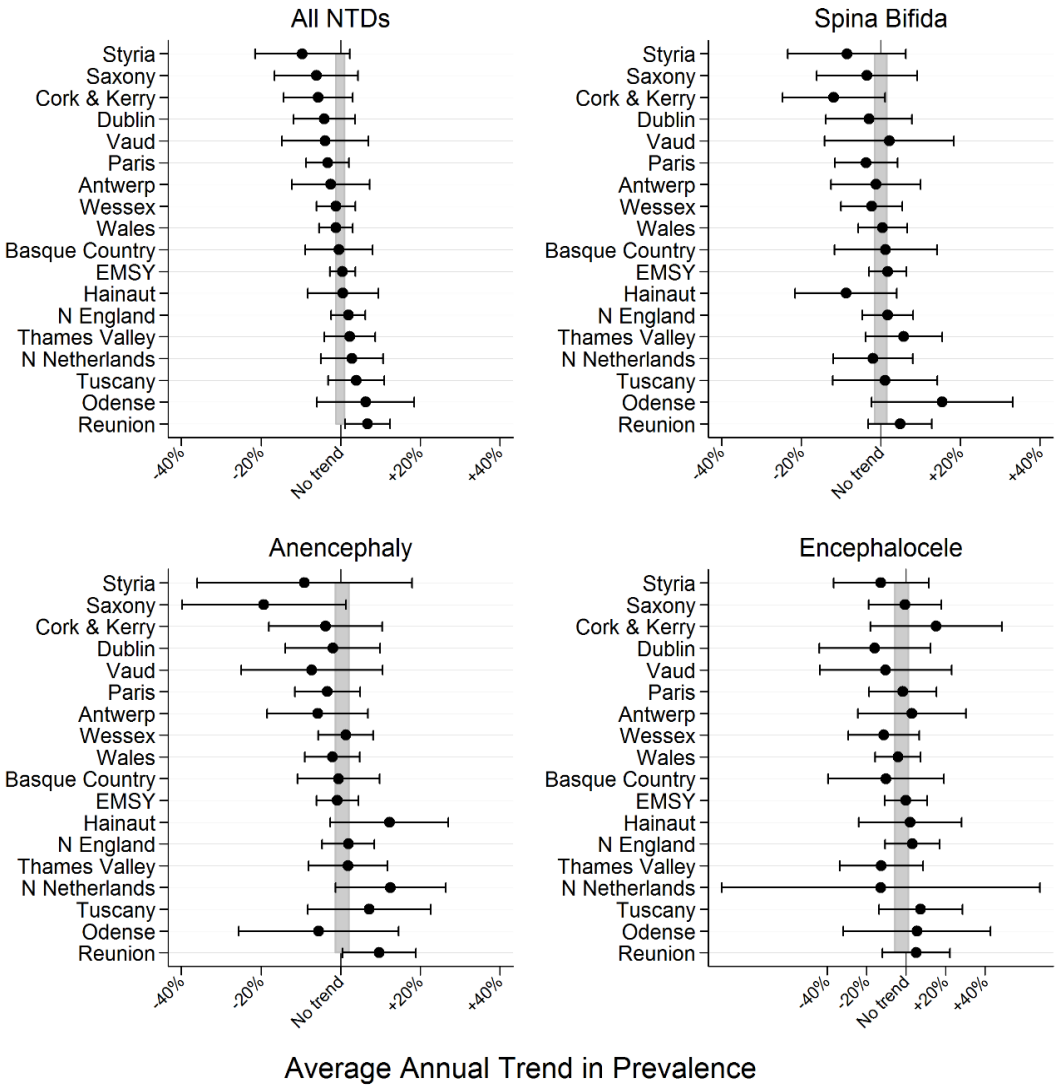


Figure A21. Individual registry results for average annual trend in prevalence of Neural Tube Defect subgroups 2003-2012, with 99% confidence band for overall estimate across all registries (grey shaded band).

Table A13. Run times for JAGS and Stan models 5, 6a and 6b.

Model	Programme	Run time in seconds	Run time in minutes
5	JAGS	49.7	0.8
	Stan	136.3	2.3
6a	JAGS	859.2	14.3
	Stan	31596.4	526.6
6b	JAGS	889.8	14.8
	Stan	4362.2	72.7

Table A14. Comparison of posterior distributions for JAGS and Stan models 5, 6a and 6b.

	JAGS			Stan		
Parameter	Mean (95% CI)	ESS	PSRF [a]	Mean (95% CI)	ESS	Rhat [a]
MODEL 5						
u0[1]	-7.78 (-7.84, -7.73)	60,545	1	-7.78 (-7.83, -7.73)	382	1.02
u0[2]	-9.16 (-9.26, -9.07)	58,569	1	-9.17 (-9.26, -9.08)	29	1.07
u0[3]	-7.61 (-7.66, -7.56)	60,046	1	-7.61 (-7.66, -7.57)	264	1.02
mu.u0	-8.16 (-11.69, -4.5)	60,844	1	-1.42 (-57.49, 58.6)	96	1.03
sigma.u0	2.28 (0.53, 7.97)	28,850	1	7.59 (4.69, 9.83)	86	1.04
u1[1]	0.002 (-0.01, 0.02)	40,700	1	0.001 (-0.01, 0.02)	37	1.07
u1[2]	-0.01 (-0.05, 0.01)	22,144	1	-0.01 (-0.05, 0.01)	77	1.04
u1[3]	-0.001 (-0.02, 0.02)	45,818	1	-0.001 (-0.02, 0.01)	202	1.01
mu.u1	-0.004 (-0.12, 0.1)	80,276	1.18	-0.03 (-0.15, 0.07)	2,299	1.01
sigma.u1	0.07 (0.001, 0.51)	471	1.2	0.083 (0.001, 0.47)	537	1.01
sigma.e	0.03 (0.001, 0.08)	2,407	1	0.026 (0.004, 0.08)	25	1.08
MODEL 6a						
u0[1]	-4.87 (-38.7, 27.96)	12	4.93	0.08 (-5.87, 3.74)	178	1.01
u0[2]	-6.24 (-40.07, 26.58)	12	4.93	-1.29 (-7.24, 2.37)	178	1.01
u0[3]	-4.68 (-38.53, 28.14)	12	4.93	0.26 (-5.68, 3.91)	178	1.01
mu.u0	-5.25 (-39.17, 27.68)	13	4.83	0.47 (-62.28, 63)	1,664	1.00
sigma.u0	2.31 (0.52, 8.03)	27,661	1	2.44 (0.5, 8.35)	249	1.01
u1[1]	0.002 (-0.02, 0.02)	30,183	1	0.002 (-0.01, 0.02)	1,409	1.00
u1[2]	-0.01 (-0.05, 0.01)	16,141	1	-0.01 (-0.05, 0.01)	665	1.01
u1[3]	-0.002 (-0.02, 0.01)	32,578	1	-0.003 (-0.02, 0.01)	3,837	1.00
mu.u1	-0.01 (-0.11, 0.09)	101,157	1.16	-0.01 (-0.12, 0.1)	1,909	1.00
sigma.u1	0.05 (0, 0.37)	2,907	1.13	0.08 (0, 0.47)	2,598	1.00
r[1]	-3.14 (-35.98, 30.7)	12	4.93	-8.09 (-11.74, -2.12)	178	1.01
r[2]	-2.9 (-35.72, 30.93)	12	4.93	-7.84 (-11.52, -1.88)	178	1.01
r[3]	-2.88 (-35.72, 30.95)	12	4.93	-7.83 (-11.48, -1.89)	178	1.01
r[4]	-3.49 (-36.3, 30.36)	12	4.93	-8.44 (-12.1, -2.48)	178	1.01
r[5]	-3.46 (-36.29, 30.39)	12	4.93	-8.41 (-12.07, -2.46)	178	1.01
r[6]	-3.26 (-36.07, 30.57)	12	4.93	-8.2 (-11.85, -2.26)	178	1.01
r[7]	-3.04 (-35.87, 30.81)	12	4.93	-7.98 (-11.66, -2.05)	178	1.01

r[8]	-3.22 (-36.07, 30.6)	12	4.93	-8.17 (-11.82, -2.21)	178	1.01
r[9]	-2.97 (-35.79, 30.88)	12	4.93	-7.92 (-11.57, -1.97)	178	1.01
r[10]	-3.08 (-35.89, 30.74)	12	4.93	-8.02 (-11.68, -2.07)	178	1.01
r[11]	-3.34 (-36.17, 30.50)	12	4.93	-8.29 (-11.96, -2.35)	178	1.01
r[12]	-2.84 (-35.66, 30.99)	12	4.93	-7.79 (-11.43, -1.82)	178	1.01
r[13]	-2.68 (-35.51, 31.16)	12	4.93	-7.62 (-11.27, -1.68)	178	1.01
r[14]	-2.41 (-35.25, 31.44)	12	4.93	-7.36 (-11.02, -1.41)	178	1.01
r[15]	-2.92 (-35.74, 30.91)	12	4.93	-7.86 (-11.53, -1.9)	178	1.01
r[16]	-2.83 (-35.66, 31.01)	12	4.93	-7.77 (-11.41, -1.83)	178	1.01
r[17]	-2.84 (-35.67, 30.99)	12	4.93	-7.79 (-11.44, -1.84)	178	1.01
r[18]	-2.66 (-35.49, 31.19)	12	4.93	-7.61 (-11.26, -1.66)	178	1.01
mu.r	-3.00 (-35.84, 30.84)	12	4.93	-7.94 (-11.6, -1.98)	178	1.01
sigma.r	0.32 (0.22, 0.48)	49,263	1	0.32 (0.22, 0.47)	2,616	1.00
sigma.e	0.12 (0.04, 0.19)	966	1.02	0.13 (0.05, 0.2)	329	1.00
MODEL 6b						
u1[1]	0.002 (-0.01, 0.02)	31,311	1	0.002 (-0.01, 0.02)	4,143	1.00
u1[2]	-0.01 (-0.05, 0.01)	15,622	1	-0.01 (-0.05, 0.01)	2,769	1.00
u1[3]	-0.001 (-0.02, 0.01)	32,375	1	-0.001 (-0.02, 0.01)	2,459	1.00
mu.u1	-0.004 (-0.1, 0.09)	97,477	1.08	-0.01 (-0.11, 0.1)	2,979	1.00
sigma.u1	0.06 (0.001, 0.39)	596	1.11	0.07 (0.002, 0.43)	5,429	1.00
u0[1], r[1]	-8.26 (-8.53, -8.00)	32,262	1	-8.26 (-8.53, -8)	14,883	1.00
u0[1], r[2]	-9.63 (-10.14, -9.19)	19,185	1	-9.63 (-10.13, -9.19)	10,957	1.00
u0[1], r[3]	-7.72 (-7.92, -7.53)	38,498	1	-7.72 (-7.92, -7.52)	21,116	1.00
u0[1], r[4]	-7.61 (-7.82, -7.41)	38,078	1	-7.61 (-7.82, -7.41)	8,552	1.00
u0[1], r[5]	-9.52 (-10.01, -9.09)	20,311	1	-9.53 (-10.03, -9.08)	3,959	1.00
u0[1], r[6]	-7.78 (-7.99, -7.58)	40,597	1	-7.78 (-8, -7.57)	4,031	1.00
u0[1], r[7]	-7.67 (-7.96, -7.4)	32,076	1	-7.67 (-7.96, -7.4)	23,890	1.00
u0[1], r[8]	-9.15 (-9.66, -8.67)	20,885	1	-9.15 (-9.66, -8.68)	8,256	1.00
u0[1], r[9]	-7.57 (-7.82, -7.33)	36,732	1	-7.57 (-7.82, -7.32)	19,012	1.00
u0[1], r[10]	-8.38 (-8.64, -8.14)	33,252	1	-8.38 (-8.64, -8.14)	4,617	1.00
u0[1], r[11]	-9.53 (-9.96, -9.15)	22,416	1	-9.52 (-9.95, -9.14)	10,257	1.00
u0[1], r[12]	-8.07 (-8.29, -7.87)	36,793	1	-8.07 (-8.3, -7.86)	3,277	1.00
u0[1], r[13]	-7.64 (-7.75, -7.52)	57,354	1	-7.64 (-7.75, -7.52)	23,866	1.00
u0[1], r[14]	-9.23 (-9.46, -9.01)	35,039	1	-9.23 (-9.46, -9)	6,919	1.00
u0[1], r[15]	-7.55 (-7.66, -7.45)	59,383	1	-7.55 (-7.67, -7.44)	4,659	1.00
u0[1], r[16]	-8.06 (-8.37, -7.78)	29,229	1	-8.06 (-8.37, -7.77)	8,423	1.00
u0[1], r[17]	-9.27 (-9.76, -8.82)	19,427	1	-9.27 (-9.78, -8.82)	2,722	1.00
u0[1], r[18]	-7.79 (-8.04, -7.55)	35,171	1	-7.79 (-8.05, -7.55)	17,260	1.00
mu.u0[1]	-7.93 (-8.17, -7.7)	52,346	1	-7.93 (-8.17, -7.7)	13,947	1.00
sigma.u0[1]	0.48 (0.32, 0.71)	37,986	1	0.48 (0.32, 0.71)	6,974	1.00
u0[2], r[1]	-7.47 (-7.62, -7.32)	48,266	1	-7.47 (-7.62, -7.32)	13,795	1.00
u0[2], r[2]	-8.99 (-9.29, -8.71)	27,881	1	-8.99 (-9.28, -8.71)	5,218	1.00
u0[2], r[3]	-7.39 (-7.54, -7.25)	49,499	1	-7.39 (-7.54, -7.25)	20,982	1.00
u0[2], r[4]	-8.24 (-8.52, -7.97)	31,910	1	-8.24 (-8.53, -7.97)	2,473	1.00
u0[2], r[5]	-9.99 (-10.68, -9.43)	12,221	1	-9.99 (-10.68, -9.43)	7,849	1.00

u0[2], r[6]	-7.72 (-7.93, -7.52)	37,689	1	-7.72 (-7.93, -7.52)	5,224	1.00
u0[2], r[7]	-7.85 (-8.26, -7.47)	21,073	1	-7.85 (-8.26, -7.46)	8,201	1.00
u0[2], r[8]	-9.17 (-9.79, -8.6)	15,916	1	-9.17 (-9.78, -8.6)	4,338	1.00
u0[2], r[9]	-7.55 (-7.87, -7.25)	30,816	1	-7.56 (-7.89, -7.25)	2,091	1.00
u0[2], r[10]	-7.71 (-7.9, -7.54)	40,529	1	-7.71 (-7.9, -7.54)	18,568	1.00
u0[2], r[11]	-9.05 (-9.38, -8.73)	26,609	1	-9.05 (-9.38, -8.73)	6,201	1.00
u0[2], r[12]	-7.63 (-7.8, -7.46)	41,043	1	-7.63 (-7.8, -7.46)	19,581	1.00
u0[2], r[13]	-7.33 (-7.54, -7.13)	38,087	1	-7.33 (-7.54, -7.13)	15,431	1.00
u0[2], r[14]	-8.71 (-9.09, -8.35)	20,234	1	-8.71 (-9.08, -8.35)	19,563	1.00
u0[2], r[15]	-7.11 (-7.29, -6.93)	37,885	1	-7.11 (-7.29, -6.93)	9,322	1.00
u0[2], r[16]	-8.5 (-8.83, -8.19)	29,356	1	-8.49 (-8.82, -8.19)	21,907	1.00
u0[2], r[17]	-8.97 (-9.35, -8.6)	23,427	1	-8.97 (-9.35, -8.61)	19,295	1.00
u0[2], r[18]	-7.54 (-7.74, -7.35)	41,382	1	-7.54 (-7.74, -7.35)	13,385	1.00
mu.u0[2]	-9.21 (-9.45, -8.98)	34,870	1	-9.21 (-9.47, -8.98)	2,579	1.00
sigma.u0[2]	0.43 (0.24, 0.69)	16,674	1	0.43 (0.25, 0.69)	21,921	1.00
u0[3], r[1]	-8.75 (-9.23, -8.32)	20,257	1	-8.74 (-9.23, -8.32)	15,226	1.00
u0[3], r[2]	-8.97 (-9.44, -8.53)	20,436	1	-8.98 (-9.43, -8.53)	12,481	1.00
u0[3], r[3]	-7.85 (-8.13, -7.59)	31,753	1	-7.85 (-8.13, -7.59)	12,676	1.00
u0[3], r[4]	-7.69 (-7.88, -7.5)	40,919	1	-7.69 (-7.88, -7.51)	12,530	1.00
u0[3], r[5]	-9.26 (-9.64, -8.92)	23,752	1	-9.26 (-9.64, -8.91)	5,839	1.00
u0[3], r[6]	-7.66 (-7.83, -7.49)	42,033	1	-7.66 (-7.84, -7.48)	12,207	1.00
u0[3], r[7]	-8.51 (-8.77, -8.27)	36,034	1	-8.52 (-8.78, -8.27)	4,559	1.00
u0[3], r[8]	-9.57 (-9.99, -9.20)	21,778	1	-9.57 (-9.99, -9.20)	16,125	1.00
u0[3], r[9]	-8.03 (-8.24, -7.84)	37,998	1	-8.03 (-8.23, -7.84)	16,010	1.00
u0[3], r[10]	-7.99 (-8.36, -7.65)	23,423	1	-7.99 (-8.36, -7.65)	4,937	1.00
u0[3], r[11]	-8.94 (-9.44, -8.45)	19,158	1	-8.95 (-9.44, -8.46)	17,097	1.00
u0[3], r[12]	-7.77 (-8.07, -7.49)	28,145	1	-7.77 (-8.07, -7.49)	12,571	1.00
u0[3], r[13]	-7.62 (-7.78, -7.47)	42,596	1	-7.62 (-7.78, -7.47)	9,952	1.00
u0[3], r[14]	-8.69 (-8.95, -8.44)	31,516	1	-8.69 (-8.95, -8.44)	16,491	1.00
u0[3], r[15]	-7.39 (-7.53, -7.26)	50,857	1	-7.39 (-7.53, -7.26)	21,386	1.00
u0[3], r[16]	-7.47 (-7.63, -7.31)	42,301	1	-7.47 (-7.63, -7.31)	13,245	1.00
u0[3], r[17]	-9.31 (-9.67, -8.98)	25,513	1	-9.32 (-9.68, -8.97)	5,350	1.00
u0[3], r[18]	-7.68 (-7.85, -7.52)	41,675	1	-7.68 (-7.85, -7.51)	2,509	1.00
mu.u0[3]	-7.66 (-7.81, -7.52)	51,122	1	-7.66 (-7.82, -7.52)	7,705	1.00
sigma.u0[3]	0.27 (0.18, 0.41)	41,844	1	0.27 (0.18, 0.41)	6,904	1.00
mu.ar [b]	-8.24 (-11.82, -4.51)	61,471	1	-8.25 (-11.95, -4.57)	5,772	1.00
sigma.a [b]	2.27 (0.50, 8.01)	30,373	1	2.27 (0.50, 8.01)	8,671	1.00
sigma.e	0.06 (0.002, 0.14)	240	1.02	0.08 (0.02, 0.15)	239	1.01

[a] PSRF and Rhat are the Estimated potential scale reduction factor for JAGs and Stan, respectively

*[b] mu.ar is the mean intercept across all CA*registry combinations, and sigma.ar the SD of the intercepts across all CA*registry combinations*

Table A15. Comparison of posterior distributions with different values for prior distributions, NTDs model 5.

Parameter	Mean (95% CI)	ESS	Mean (95% CI)	ESS	Mean (95% CI)	ESS
Prior for variances / means	*Uniform(0, 10) / Normal(0, 0.001)*		*Uniform(0, 100) / Normal(0, 0.001)*		*Uniform(0, 10) / Normal (0, 0.0001)*	
u0[1]	-7.784 (-7.837, -7.731)	60,545	-7.784 (-7.837, -7.731)	59,981	-7.784 (-7.837, -7.731)	60,000
u0[2]	-9.162 (-9.261, -9.065)	58,569	-9.162 (-9.262, -9.065)	59,112	-9.162 (-9.26, -9.066)	57,162
u0[3]	-7.612 (-7.662, -7.563)	60,046	-7.612 (-7.662, -7.563)	60,000	-7.612 (-7.661, -7.563)	60,474
mu.u0	-8.16 (-11.69, -4.499)	60,844	-8.058 (-14.054, -1.275)	28,326	-8.174 (-11.92, -4.39)	59,079
sigma.u0	2.282 (0.525, 7.969)	28,850	3.846 (0.529, 21.678)	1,639	2.308 (0.524, 8.026)	28,557
u1[1]	0.002 (-0.014, 0.02)	40,700	0.002 (-0.014, 0.02)	42,805	0.002 (-0.014, 0.02)	42,921
u1[2]	-0.012 (-0.046, 0.013)	22,144	-0.012 (-0.046, 0.013)	24,708	-0.012 (-0.046, 0.013)	26,008
u1[3]	-0.001 (-0.017, 0.015)	45,818	-0.001 (-0.017, 0.015)	46,148	-0.001 (-0.017, 0.015)	48,094
mu.u1	-0.004 (-0.118, 0.102)	80,276	-0.003 (-0.105, 0.095)	70,369	-0.004 (-0.1, 0.085)	61,964
sigma.u1	0.07 (0.001, 0.508)	471.5	0.058 (0.001, 0.41)	736	0.049 (0.001, 0.332)	1,852
sigma.e	0.029 (0.001, 0.079)	2,407	0.03 (0.001, 0.079)	2,476	0.03 (0.002, 0.078)	2,602
Prior for variances / means	*Cauchy, scale=5 / Normal(0, 0.001)*		*Cauchy, scale=10 / Normal(0, 0.001)*		*Cauchy, scale=25 / Normal(0, 0.0001)*	
u0[1]	-7.784 (-7.837, -7.732)	60,000	-7.784 (-7.837, -7.731)	60,014	-7.784 (-7.837, -7.732)	59,319
u0[2]	-9.162 (-9.261, -9.066)	60,633	-9.162 (-9.261, -9.066)	59,227	-9.162 (-9.26, -9.066)	60,804
u0[3]	-7.612 (-7.662, -7.564)	60,652	-7.612 (-7.662, -7.563)	60,012	-7.612 (-7.662, -7.563)	60,875
mu.u0	-8.167 (-11.12, -5.168)	55,584	-8.139 (-11.76, -4.198)	64,587	-8.146 (-12.9, -3.157)	57,867
sigma.u0	1.976 (0.508, 7.035)	11,925	2.426 (0.52, 9.615)	6,151	2.986 (0.53, 13.621)	6,244
u1[1]	0.002 (-0.014, 0.02)	42,484	0.002 (-0.014, 0.02)	45,339	0.002 (-0.014, 0.02)	42,277
u1[2]	-0.012 (-0.046, 0.013)	25,292	-0.012 (-0.046, 0.013)	22,904	-0.012 (-0.047, 0.013)	18,831
u1[3]	-0.001 (-0.017, 0.015)	45,519	-0.001 (-0.017, 0.015)	47,543	-0.001 (-0.017, 0.015)	46,176
mu.u1	-0.004 (-0.106, 0.09)	45,590	-0.003 (-0.111, 0.104)	112,730	-0.002 (-0.184, 0.18)	82,230
sigma.u1	0.057 (0.001, 0.401)	1,093	0.08 (0.001, 0.474)	836	0.126 (0.001, 1.462)	315
sigma.e	0.03 (0.002, 0.078)	2,142	0.03 (0.001, 0.078)	2,663	0.03 (0.001, 0.078)	2,581

A5. Supplementary results for analysis of prevalence in chromosomal anomalies

Table A16. Model fit for hierarchical models including five chromosomal CA subgroups together.

Model	***3**: Frequentist hierarchical model*	***4**: Frequentist hierarchical model + registry*	***5**: Bayesian hierarchical model*	***6A**: Bayesian hierarchical model + registry (A)*	***6B**: Bayesian hierarchical model + registry (B)*
cAIC	422	4465	-	-	-
Deviance	456	4565	-	-	-
Residual DF	44	888	-	-	-
Mean Deviance	-	-	400	4333	4246
Penalty	-	-	20	130	136
DIC (penalised deviance)	-	-	420	4463	4382
Multivariate PSRF	-	-	1.00	2.54	1.01
Mean SD for overdispersion parameter	0.035	0.104	0.037	0.103	0.066
Mean SD of registry intercepts	-	0.272	-	0.296	1.941 [a]
Mean SD of CA intercepts	1.098	1.102	1.906	1.915	
Mean SD of CA trends	0.026	0.026	0.054	0.054	0.050
Estimated correlation between CA intercepts and slopes	0.820	0.830	-	-	-

*[a] SD of the intercepts across all CA*registry combinations*

Table A17. Model fit for hierarchical models including three chromosomal trisomy subgroups.

Model	***3**:*	***4**:*	***5**:*	***6A**:*	***6B**:*

	Frequentist hierarchical model	*Frequentist hierarchical model + registry*	*Bayesian hierarchical model*	*Bayesian hierarchical model + registry (A)*	*Bayesian hierarchical model + registry (B)*
cAIC	299	2998	-	-	-
Deviance	293	3077	-	-	-
Residual DF	175	530	-	-	-
Mean Deviance	-	-	260	2906	2885
Penalty	-	-	15	95	93
DIC (penalised deviance)	-	-	275	3001	2978
Multivariate PSRF	-	-	1.00	2.11	1.01
Mean SD for overdispersion parameter	0.034	0.091	0.041	0.090	0.063
Mean SD of registry intercepts	-	0.272	-	0.295	2.858[a]
Mean SD of CA intercepts	0.961	0.966	2.855	2.858	
Mean SD of CA trends	0.0003	0.004	0.032	0.016	0.033
Estimated correlation between CA intercepts and slopes	1	1	-	-	-

[a] *SD of the intercepts across all CA*registry combinations*

A6. Supplementary results for analysis of prevalence in digestive system CAs

Table A18. Model fit for hierarchical models including digestive system CA subgroups.

Model	***3**: Frequentist hierarchical model*	***4**: Frequentist hierarchical model + registry*	***5**: Bayesian hierarchical model*	***6A**: Bayesian hierarchical model + registry (A)*	***6B**: Bayesian hierarchical model + registry (B)*
cAIC	599	5009	-	-	-
Deviance	589	5067	-	-	-
Residual DF	74	1425	-	-	-
Mean Deviance	-	-	523	4894	4713
Penalty	-	-	29	158	123
DIC (penalised deviance)	-	-	552	5007	4836
Multivariate PSRF	-	-	1.00	2.55	1.03
Mean SD for overdispersion parameter	0.075	0.190	0.075	0.187	0.058
Mean SD of registry intercepts	-	0.141	-	0.153	1.382 [a]
Mean SD of CA intercepts	0.991	0.991	1.320	1.322	
Mean SD of CA trends	0.014	0.015	0.018	0.018	0.018
Estimated correlation between CA intercepts and slopes	0.99	0.97	-	-	-

[a] *SD of the intercepts across all CA*registry combinations*

A7. Supplementary results for analysis of prevalence in congenital heart defects

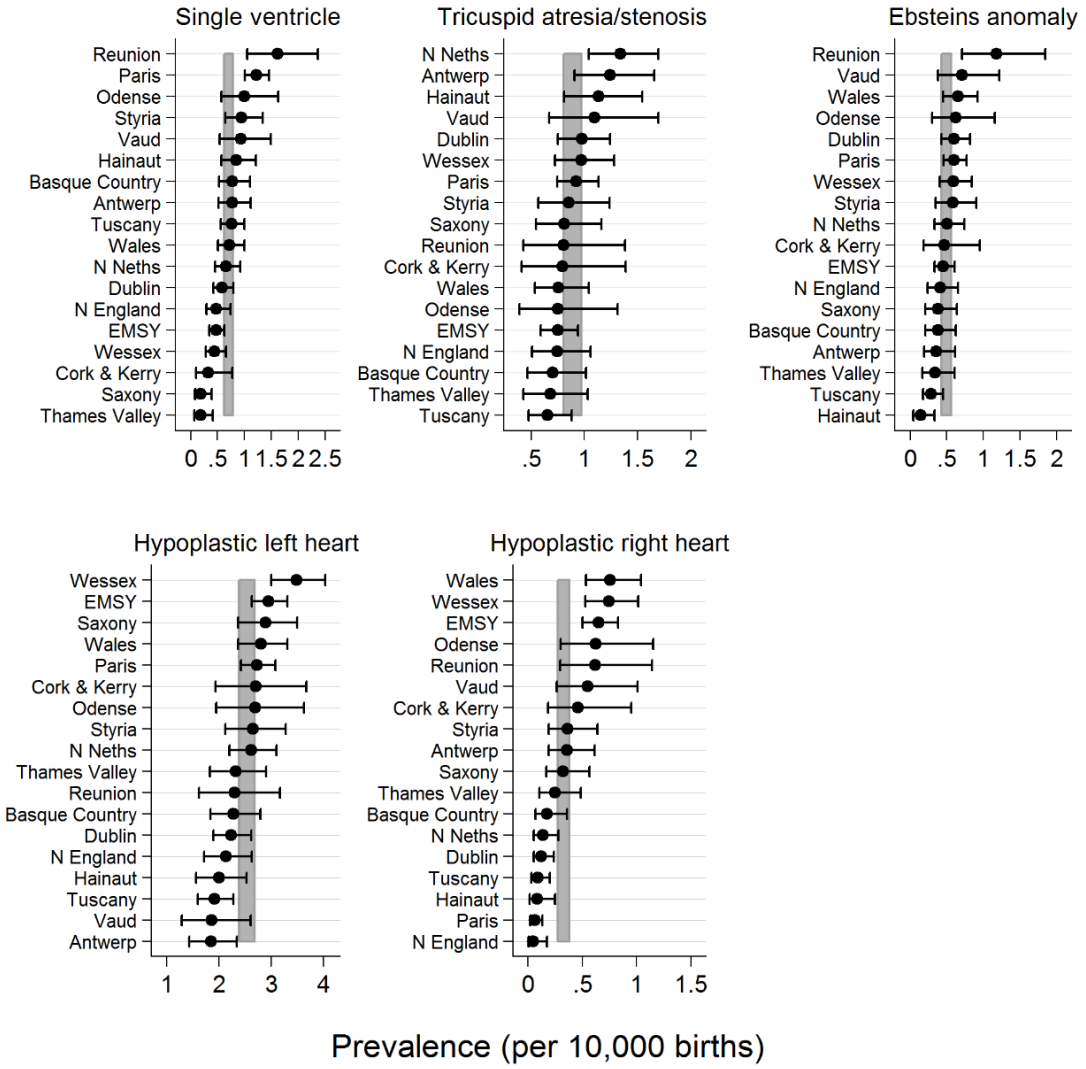


Figure A22. Total prevalence of congenital heart defect subgroups in severity group 1 for 18 EUROCAT registries from 2003 to 2012.

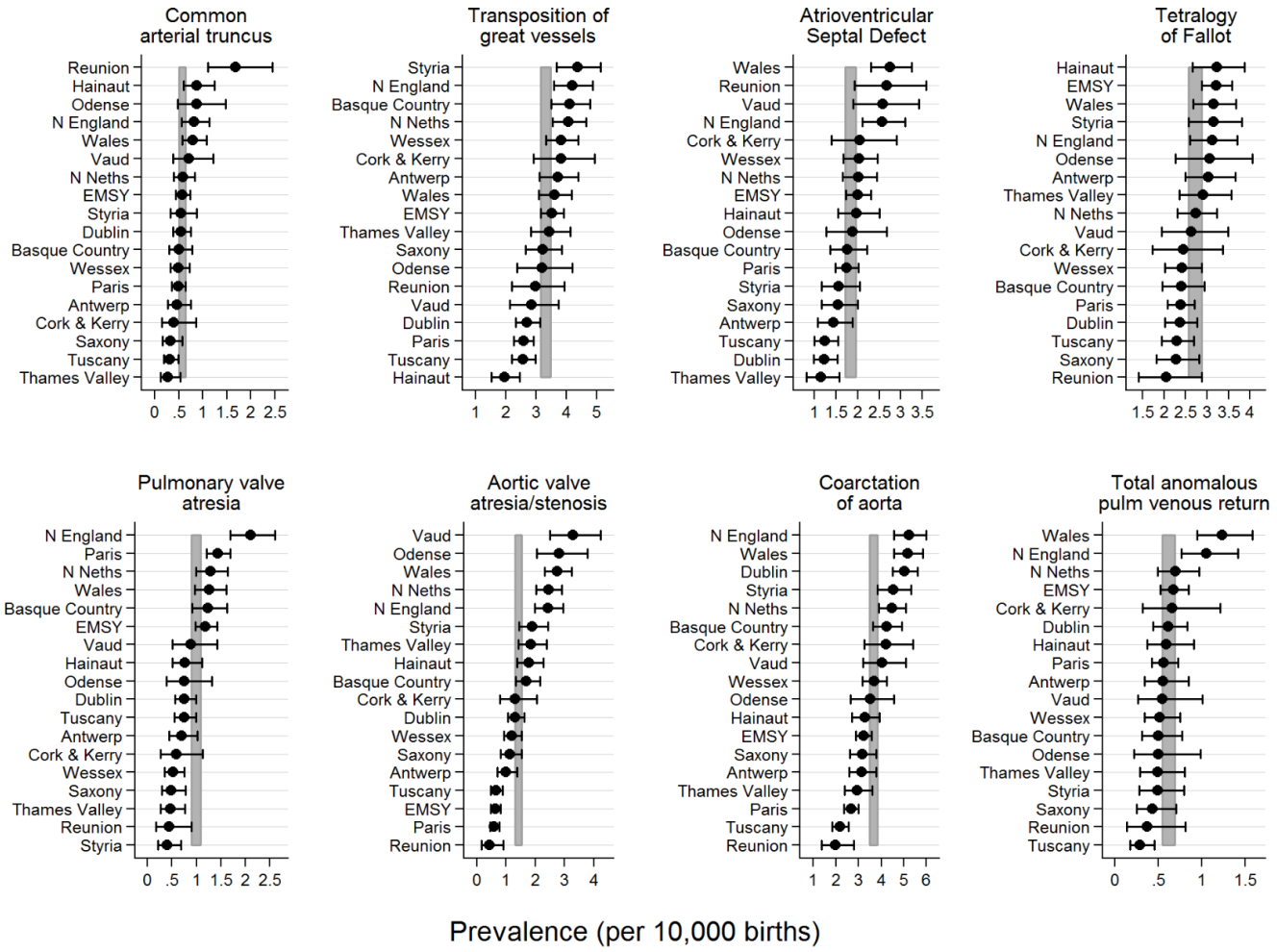


Figure A23. Total prevalence of congenital heart defect subgroups in severity group 2 for 18 EUROCAT registries from 2003 to 2012.

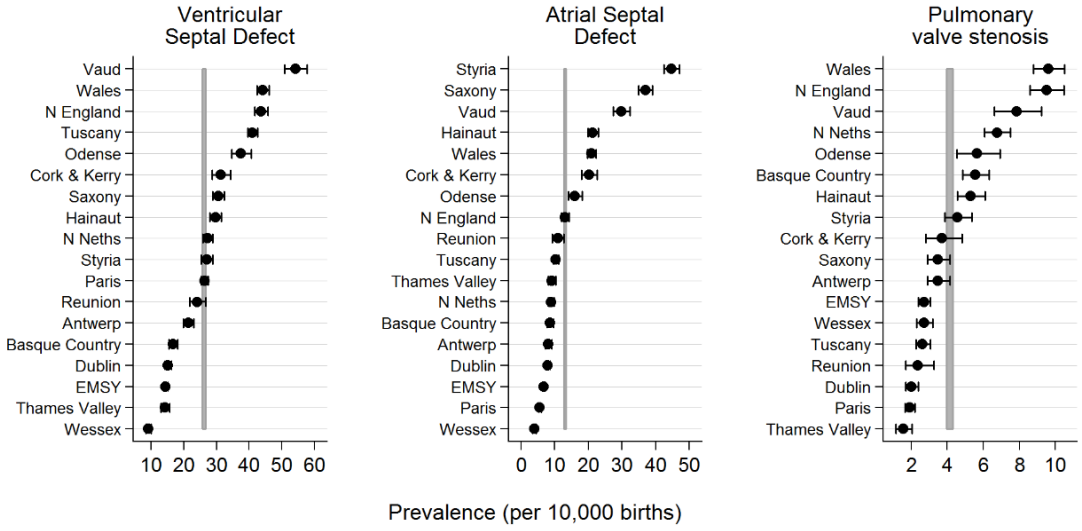


Figure A24. Total prevalence of congenital heart defect subgroups in severity group 3 for 18 EUROCAT registries from 2003 to 2012.

Table A19. Model 5 with additional severity grouping; summary of posterior distribution for overall and severity group-level parameters.

Type	Parameter	Mean	SD	2.5%	50%	97.5%	ESS
Mean intercepts	mu.u0	-8.482	1.989	-12.530	-8.637	-3.874	54,917
	mu.u0.grp[1]	-9.330	0.545	-10.350	-9.354	-8.180	45,378
	mu.u0.grp[2]	-8.747	0.328	-9.403	-8.748	-8.089	54,727
	mu.u0.grp[3]	-7.508	1.372	-10.180	-7.342	-4.990	33,926
Mean trends	mu.u1	-0.001	0.115	-0.150	0.000	0.147	79,848
	mu.u1.grp[1]	0.004	0.017	-0.030	0.003	0.038	43,386
	mu.u1.grp[2]	0.006	0.008	-0.011	0.006	0.022	26,452
	mu.u1.grp[3]	-0.012	0.038	-0.067	-0.012	0.042	47,969
SD for intercepts	sigma.u0	2.554	2.184	0.134	1.863	8.522	24,626
	sigma.u0.grp[1]	1.121	0.680	0.468	0.936	2.917	27,709
	sigma.u0.grp[2]	0.892	0.318	0.498	0.822	1.710	45,627
	sigma.u0.grp[3]	2.528	1.914	0.584	1.899	7.951	32,027
SD for trends	sigma.u1	0.079	0.178	0.001	0.028	0.526	1,706
	sigma.u1.grp[1]	0.033	0.030	0.002	0.026	0.103	8,709
	sigma.u1.grp[2]	0.014	0.010	0.001	0.013	0.037	10,593
	sigma.u1.grp[3]	0.067	0.145	0.006	0.035	0.302	1,266
Overdispersion	sigma.e	0.083	0.012	0.062	0.082	0.107	18,183
Severity group 1:							
Single ventricle	u1[3]	0.023	0.022	-0.013	0.020	0.069	22,140
Tricuspid atresia and stenosis	u1[8]	0.013	0.018	-0.020	0.012	0.053	35,006
Ebsteins anomaly	u1[9]	-0.019	0.023	-0.070	-0.016	0.019	23,349
Hypoplastic left heart	u1[13]	-0.006	0.013	-0.032	-0.005	0.018	35,980
Hypoplastic right heart	u1[14]	0.012	0.020	-0.025	0.010	0.056	36,733
Severity group 2:							
Common arterial truncus	u1[1]	1.23E-08	0.015	-0.034	0.002	0.025	28,137
Transposition of great vessels	u1[2]	0.007	0.010	-0.012	0.007	0.027	43,341
Atrioventricular Septal Defect	u1[6]	0.013	0.011	-0.007	0.012	0.038	29,400
Tetralogy of Fallot	u1[7]	0.019	0.012	-0.002	0.017	0.045	18,907
Pulmonary valve atresia	u1[11]	0.004	0.012	-0.022	0.005	0.027	38,088
Aortic valve atresia/stenosis	u1[12]	0.006	0.011	-0.017	0.007	0.029	42,959
Coarctation of aorta	u1[15]	-0.003	0.011	-0.027	-0.002	0.016	19,583
Total anomalous pulmonary venous return	u1[16]	0.007	0.013	-0.019	0.007	0.034	43,425

Severity group 3:							
Ventricular Septal Defect	u1[4]	-0.004	0.010	-0.024	-0.003	0.016	45,651
Atrial Septal Defect	u1[5]	-0.038	0.011	-0.058	-0.038	-0.017	48,068
Pulmonary valve stenosis	u1[10]	-0.026	0.012	-0.049	-0.026	-0.004	53,236

Table A20. Model fit for hierarchical models including congenital heart defect subgroups.

Model:	***3**: Frequentist hierarchical model*	***4**: Frequentist hierarchical model + registry*	***5**: Bayesian hierarchical model*	***5sev**: Bayesian hierarchical model + severity [a]*	***6A**: Bayesian hierarchical model + registry (A)*	***6B**: Bayesian hierarchical model + registry (B)*
cAIC	1311	13,189	-	-	-	-
Deviance	1451	14,277	-	-	-	-
Residual DF	165	3036	-	-	-	-
Mean Deviance	-	-	1226	1149	11,128	11,857
Penalty	-	-	85	75	1192	856
DIC (penalised deviance)	-	-	1312	1224	12,320	12,722
Multivariate PSRF	-	-	1.00	1.00	1.25	1.02
Mean SD for overdispersion parameter	0.090	0.466	0.092	0.083	0.439	0.247
Mean SD of registry intercepts	-	0.283	-	-	0.288	1.277 [b]
Mean SD of CA intercepts	1.122	1.132	1.260	2.553	1.315	
Mean SD of CA trends	0.017	0.006	0.019	0.079	0.011	0.015
Estimated correlation between CA intercepts and slopes	-0.47	-0.89	-	-	-	-

[a] *Model 5 including random effects for severity group*

[b] *SD of the intercepts across all CA*registry combinations*

A8. Cavadino A et al (2016). Use of hierarchical models to analyze European trends in congenital anomaly prevalence. Birth Defects Res A Clin Mol Teratol, 106, 480-8.

Use of Hierarchical Models to Analyze European Trends in Congenital Anomaly Prevalence

Alana Cavadino[1], David Prieto-Merino[2,3,4], Marie-Claude Addor[5], Larraitz Arriola[6], Fabrizio Bianchi[7], Elizabeth Draper[8], Ester Garne[9], Ruth Greenlees[10], Martin Haeusler[11], Babak Khoshnood[12], Jenny Kurinczuk[13], Bob McDonnell[14], Vera Nelen[15], Mary O'Mahony[16], Hanitra Randrianaivo[17], Judith Rankin[18], Anke Rissmann[19], David Tucker[20], Christine Verellen-Dumoulin[21], Hermien de Walle[22], Diana Wellesley[23], and Joan K. Morris*[1]

Background: Surveillance of congenital anomalies is important to identify potential teratogens. Despite known associations between different anomalies, current surveillance methods examine trends within each subgroup separately. We aimed to evaluate whether hierarchical statistical methods that combine information from several subgroups simultaneously would enhance current surveillance methods using data collected by EUROCAT, a European network of population-based congenital anomaly registries. **Methods:** Ten-year trends (2003 to 2012) in 18 EUROCAT registries over 11 countries were analyzed for the following groups of anomalies: neural tube defects, congenital heart defects, digestive system, and chromosomal anomalies. Hierarchical Poisson regression models that combined related subgroups together according to EUROCAT's hierarchy of subgroup coding were applied. Results from hierarchical models were compared with those from Poisson models that consider each congenital anomaly separately. **Results:** Hierarchical models gave similar results as those obtained when considering each anomaly subgroup in a separate analysis. Hierarchical models that included only around three subgroups showed poor convergence and were generally found to be over-parameterized. Larger sets of anomaly subgroups were found to be too heterogeneous to group together in this way. **Conclusion:** There were no substantial differences between independent analyses of each subgroup and hierarchical models when using the EUROCAT anomaly subgroups. Considering each anomaly separately, therefore, remains an appropriate method for the detection of potential changes in prevalence by surveillance systems. Hierarchical models do, however, remain an interesting alternative method of analysis when considering the risks of specific exposures in relation to the prevalence of congenital anomalies, which could be investigated in other studies.

Birth Defects Research (Part A) 106:480–10, 2016.

Key words: congenital anomalies; trends; prevalence; hierarchical models

Introduction

Congenital anomalies are structural or functional abnormalities that are present at birth. They are a leading worldwide cause of fetal and infant death, chronic illness, and disability in childhood; a diverse group of disorders for which only around 50% can be linked to a specific known cause or risk factor (World Health Organization, 2014). Causes of congenital anomaly include a wide range of both genetic and

Additional Supporting information may be found in the online version of this article.

[1]Wolfson Institute of Preventive Medicine, Queen Mary University of London, United Kingdom
[2]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom
[3]Farr Institute of Health Informatics Research, University College London, United Kingdom
[4]Catholic University of Murcia (UCAM), Spain
[5]Service de Genetique Medicale Maternite, CHUV, Lausanne, Switzerland
[6]Public Health Division of Gipuzkoa, Instituto BIO-Donostia Basque Government CIBER Epidemiología y Salud Pública - CIBERESP, San Sebatian, Spain
[7]CNR Institute of Clinical Physiology and Tuscany Registry of Congenital Defects, "Gabriele Monasterio" Foundation, Pisa, Italy
[8]Department of Health Sciences, University of Leicester, Leicester, United Kingdom
[9]Paediatric Department, Hospital Lillebaelt-Kolding, Denmark
[10]Institute of Nursing and Health Research, Ulster University, Newtownabbey, United Kingdom
[11]Medical University of Graz, Austria
[12]Obstetrical, Perinatal and Pediatric Epidemiology Research Team, Center for Biostatistics and Epidemiology, INSERM U1153, Maternité de Port-Royal, Paris, France
[13]National Perinatal Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, United Kingdom
[14]Health Service Executive, Dublin, Ireland
[15]Provincial Institute for Hygiene, Antwerp, Belgium
[16]Department of Public Health, Health Service Executive - South, Ireland
[17]Medical Genetics Unit of CHU Sud Réunion, Ile de la Reunion, France
[18]Institute of Health & Society, Newcastle University, Newcastle, United Kingdom
[19]Malformation Monitoring Centre Saxony-Anhalt, Medical Faculty Otto-von-Guericke University, Magdeburg, Germany
[20]Public Health Wales, Swansea, United Kingdom
[21]Center for Human Genetics, Institut de Recherche Scientifique en Pathologie et en Génétique, Charleroi, Belgium
[22]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, Netherlands
[23]University Hospitals Southampton, Faculty of Medicine and Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, United Kingdom
Supported by the Medical Research Council (grant number 1504916).
*Correspondence to: Joan K. Morris, Centre for Environmental and Preventive Medicine, Wolfson Institute of Preventive Medicine, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ.E-mail: j.k.morris@qmul.ac.uk

Published online 15 June 2016 in Wiley Online Library (wileyonlinelibrary.com). Doi: 10.1002/bdra.23515

environmental factors such as maternal age, nutritional status, or exposure to certain medications. It is important to identify risk factors for congenital anomalies, in particular the early identification of new potentially teratogenic exposures.

Following the thalidomide disaster, congenital anomaly registries were established worldwide to facilitate surveillance and research into the causes of birth defects (McBride, 1961; Khoury et al., 1994). A European network of such population-based registries, EUROCAT, provides important epidemiologic information on congenital anomalies by collecting data on over 1.7 million births from 43 registries in 23 countries across Europe (EUROCAT, 2016). EUROCAT annually monitors the birth prevalence of specific anomalies to detect new or continuing trends, identifying new potentially teratogenic exposures and evaluating the effectiveness of primary prevention policies (Dolk, 2005).

Surveillance of congenital anomalies is often performed using defined sets of subgroups, such as the EUROCAT anomaly subgroups (EUROCAT, 2005). Many of these subgroups overlap, for example, the congenital heart defects (CHD) subgroup includes further subgroups such as ventricular septal defects, atrial septal defects, and tetralogy of Fallot (ToF). Despite known relationships among many of the subgroups, current surveillance methods examine trends, clusters, or associations between risk factors and anomalies within each subgroup separately (Loane et al., 2011b; EUROCAT, 2015). Relevant information on relationships between anomalies across the different subgroups is, therefore, not currently being incorporated in surveillance analyses; hence, it is possible that important associations or trends are not being detected by the current methods. Congenital anomaly surveillance methods that combine information from several subgroups simultaneously may enhance the analysis of any particular anomaly by considering what is happening in related or similar anomalies. The aim of this study is to evaluate whether hierarchical statistical methods that combine information from several subgroups within the same congenital anomaly group simultaneously increase the power to detect trends in congenital anomalies.

Materials and Methods

EUROCAT DATASET

This study is based on routinely collected EUROCAT data from 18 full member registries in 11 European countries: Austria (Styria registry), Belgium (Antwerp and Hainaut), Denmark (Odense), France (Paris and Isle de la Reunion), Germany (Saxony-Anhalt) Ireland (Cork & Kerry and Dublin), Italy (Tuscany), Netherlands (Northern Netherlands), Spain (Basque Country), Switzerland (Vaud), and the United Kingdom (East Midlands & South Yorkshire, Northern England, Thames Valley, Wales, and Wessex). Data were extracted from the EUROmediCAT central database in February 2015, including only registries with a total prevalence of all anomalies greater than 2% and available data for at least 9 years of the 10-year period from 01 January 2003 to 31 December 2012. All coding was done according to EUROCAT guide 1.3 (www.eurocat-network.eu/content/EUROCAT-Guide-1.3.pdf) (EUROCAT, 2005), which uses a hierarchy of codes to classify all cases of nonminor congenital anomaly into 89 EUROCAT anomaly subgroups. EUROCAT anomaly subgroups are grouped in a hierarchical structure, with the highest level being the major organ groups, within which there are further classes.

Spina bifida, for example, is in the neural tube defects (NTD) subgroup, which is within the nervous system group of anomalies. A case may be counted only once in each of the lowest level EUROCAT subgroups, but if it has multiple anomalies, it will be counted in multiple subgroups. Cases with genetic conditions (genetic syndromes/microdeletions, teratogenic syndromes with malformations, or chromosomal anomalies) were excluded from all analyses of nonchromosomal anomaly. Data are collected for all birth outcomes, including live and stillbirths and terminations of pregnancy for fetal anomaly. Further details regarding the registries, methods of case ascertainment, and data collection and processing are described elsewhere (EUROCAT, 2005; Boyd et al., 2011; Greenlees et al., 2011).

STATISTICAL METHODS

The most recent 10 years of data available were assessed for changes in prevalence for the following groups of anomalies: NTDs, autosomal chromosome anomalies, CHDs, and digestive system anomalies. Poisson regression was used to model prevalence rates for the number of congenital anomaly cases each year, with the log total births included as an offset to account for the differing population size each year. Estimated average yearly 10-year trends in prevalence obtained from *individual models* (separate Poisson models for each anomaly subgroup with no information sharing between anomaly subgroups) were compared with those obtained from *hierarchical models* (one Poisson model fitting related anomaly subgroups simultaneously with sharing of information between anomaly subgroups).

For CHDs, there are 16 standard subgroups (EUROCAT, 2005) that have previously been grouped using a hierarchical severity ranking according to perinatal mortality rates in nonchromosomal cases, formed of three ordered groups from severity I (high perinatal mortality) to severity III (low perinatal mortality) (EUROCAT, 2009) (Table 1). A two level hierarchy that includes the grouping of CHDs by these severity subgroups was also considered. A data-level variance component was used to directly model potential overdispersion in the data for hierarchical models (Gelman and Hill,

TABLE 1. *Total Prevalence of Selected Groups and Subgroups of Congenital Anomalies per 10,000 births, Using Data Covering 4,097,142 Births from 18 EUROCAT Registries, 2003 to 2012*

Anomaly group and subgroup	Total cases[a]	Prevalence (95% CI)
Neural tube defects	**4,167**	**10.2 (9.9, 10.5)**
Anencephaly	1,709	4.2 (4.0, 4.4)
Encephalocele	430	1.0 (1.0, 1.2)
Spina bifida	2,028	4.9 (4.7, 5.2)
Autosomal chromosome anomalies	**13,358**	**32.6 (32.1, 33.2)**
Down syndrome / trisomy 21	9,854	24.1 (23.6, 24.5)
Patau syndrome / trisomy 13	942	2.3 (2.2, 2.5)
Edwards syndrome / trisomy 18	2,562	6.3 (6.0, 6.5)
Congenital heart defects	**25,273**	**61.7 (60.9, 62.4)**
Severity group I:		
Single ventricle	249	0.6 (0.5, 0.7)
Tricuspid atresia and stenosis	286	0.7 (0.6, 0.8)
Ebstein's anomaly	212	0.5 (0.5, 0.6)
Hypoplastic left heart	1,127	2.8 (2.6, 2.9)
Hypoplastic right heart	205	0.5 (0.4, 0.6)
Severity group II:		
Common arterial truncus	233	0.6 (0.5, 0.6)
Transposition of great vessels	1,467	3.6 (3.4, 3.8)
Atrioventricular septal defect	838	2.0 (1.9, 2.2)
Tetralogy of Fallot	1,187	2.9 (2.7, 3.1)
Pulmonary valve atresia	425	1.0 (0.9, 1.1)
Aortic valve atresia/stenosis	540	1.3 (1.2, 1.4)
Coarctation of aorta	1,488	3.6 (3.4, 3.8)
Total anomalous pulmonary venous return	299	0.7 (0.6, 0.8)
Severity group III:		
Ventricular septal defect	11,262	27.5 (27.0, 28.0)
Atrial septal defect	5,226	12.8 (12.4, 13.1)
Pulmonary valve stenosis	1,850	4.5 (4.3, 4.7)
Digestive system anomalies	**7,683**	**18.8 (18.3, 19.2)**
Oesophageal atresia with or without tracheo-esophageal fistula	890	2.2 (2.0, 2.3)
Duodenal atresia or stenosis	377	0.9 (0.8, 1.0)
Atresia or stenosis of other parts of small intestine	393	1.0 (0.9, 1.1)
Ano-rectal atresia and stenosis	1,157	2.8 (2.7, 3.0)
Hirschsprung's disease	548	1.3 (1.2, 1.5)
Atresia of bile ducts	122	0.3 (0.2, 0.4)
Annular pancreas	57	0.1 (0.1, 0.2)
Diaphragmatic hernia	1,030	2.5 (2.4, 2.7)

[a]Including livebirths, stillbirths, and terminations of pregnancy after prenatal diagnosis.

2007). Models were also repeated with the inclusion of a term to take account of the random effects of registry.

All statistical analyses were conducted in R (R Development Core Team, 2008). Markov Chain Monte Carlo sampling methods were used to obtain estimates of variability around the random effects in hierarchical models by using Gibbs sampling (Casella and George, 1992) in the Bayesian analysis program JAGS by means of the R package rjags (Plummer, 2003). Results from hierarchical models are presented as annual percentage changes in prevalence and their 95% posterior credible intervals (PCIs), which can be thought of as the Bayesian equivalent of 95% confidence intervals (CIs) and where we say there is a 95% probability that the true trend in prevalence lies within this interval. If the 95% CI or PCI does not include zero then we consider this a "statistically significant" average annual change in prevalence or a "signal." The resulting estimates are only valid if convergence has occurred, which is assessed graphically and by using convergence diagnostics in the R package coda (Plummer et al., 2003). Further details on the use of the Bayesian hierarchical models in JAGs can be found in the Supplementary Materials, which are available online.

Results

A total of 103,507 cases of congenital anomaly (81,147 cases excluding genetic conditions) were available for analysis from a combined population of 4,097,142 births over 18 registries during the 10-year study period. Trends were assessed in 4167 NTD, 13,358 chromosomal, 25,273 CHD, and 7683 digestive system anomaly cases (Table 1). The rarest subgroup included in these analyses was the digestive system anomaly annular pancreas, with only 57 cases in the combined population over the 10 years giving an estimated total prevalence of 0.1 cases per 10,000 births. The most common anomaly subgroup was the CHD ventricular septal defect, with an estimated total prevalence of almost 28 cases per 10,000 births (Table 1).

MODELS FOR NTDS

There were no changes in prevalence for any of the NTD subgroups, with estimated average annual trends remaining very similar for individual and hierarchical models and 95% CIs and PCIs including zero (no change) for all estimates (Fig. 1). There was some "shrinkage" in the estimates toward the group mean in the hierarchical model, in particular for encephalocele, although this estimated trend was not significant in either model.

MODELS FOR CHROMOSOMAL ANOMALIES

In individual models, an increasing trend of 1.7% (95% CI, 0.7–2.6%) and 1.8% (95% CI, 0.4–3.1%) per year on average was observed for Down and Edwards syndromes, respectively (Fig. 2), but there was no significant change in prevalence of Patau syndrome. Trends in prevalence were similar when combining the three anomalies together

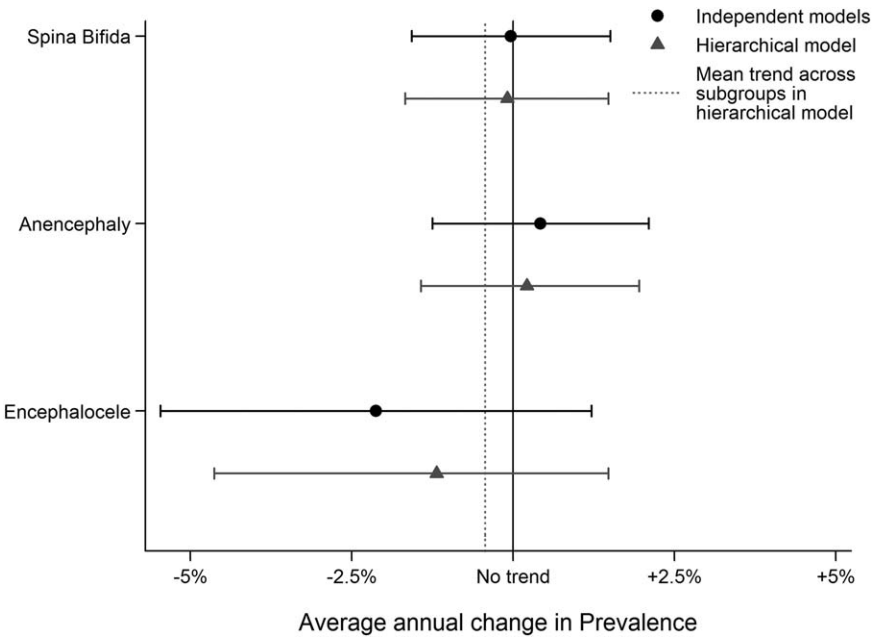


FIGURE 1. *Estimated average annual trends in prevalence of neural tube defects with 95% posterior credibility intervals.*

in a hierarchical model; the estimated trend for Patau syndrome increased slightly toward the average of the three trends but the 95% PCI still included zero (Fig. 2).

MODELS FOR CONGENITAL HEART DEFECTS

Of all cases with CHD, 85.5% were counted in one of the three EUROCAT severity groups for CHDs, excluding those with patent ductus arteriosus in term infants as well as several other CHDs that are not assigned a specific subgroup code according to EUROCAT's coding hierarchy. In individual models, decreasing trends for atrial septal defect (ASD) and pulmonary valve stenosis (PVS), and an increasing trend for ToF were observed (Fig. 3). When using a hierarchical model that combined all CHDs (Fig. 3), the estimated trends for PVS and ToF attenuated toward the null. The only significant change in prevalence in the hierarchical model was for ASD, which attenuated slightly to 3.0% on average from the estimated 4.1% in an individual model. Average annual changes in prevalence for the other CHD subgroups were a mix of increasing and decreasing trends, none of which were statistically significant in either model. When including severity subgroup as an additional level in a hierarchical model for CHDs (Fig. 3), the trends for ASD and PVS remained significant, with estimated average changes in prevalence very similar to those obtained in individual models. The increasing trend for ToF was not statistically significant when grouping all CHDs together, whether including the severity grouping or not.

MODELS FOR DIGESTIVE SYSTEM ANOMALIES

There were no significant trends in prevalence for any of the digestive system subgroups for individual or hierarchical models (Fig. 4), with estimated trends in the hierarchical model generally attenuating toward the mean of the estimated trends across the eight subgroups, which was again close to the null value of no trend. Subgroups that were less precisely estimated were more affected by the information in other subgroups, giving more marked differences in estimated trends in the less common anomalies for individual models compared with a hierarchical model.

MODEL ASSESSMENT FOR HIERARCHICAL MODELS

Parameters for hierarchical models that included a reasonable number of subgroups (i.e., eight or more for the digestive system anomalies) displayed good convergence. Hierarchical models for smaller groups of anomalies (e.g., models for NTDs and autosomal anomalies including only three subgroups) showed poor convergence due to over parameterization in the model. Further details on model diagnostics for hierarchical models are given in the Supplementary Materials.

INCLUDING A REGISTRY EFFECT

All models were repeated with the inclusion of a random effect for registry to assess the effect of accounting for differences at the registry level. The estimated trends in prevalence of each anomaly subgroup remained very similar to those described above when including the effect of registry for all models (data not shown). Hierarchical models that included a registry effect, in particular those with only a small number of subgroups, demonstrated an overall lack of convergence.

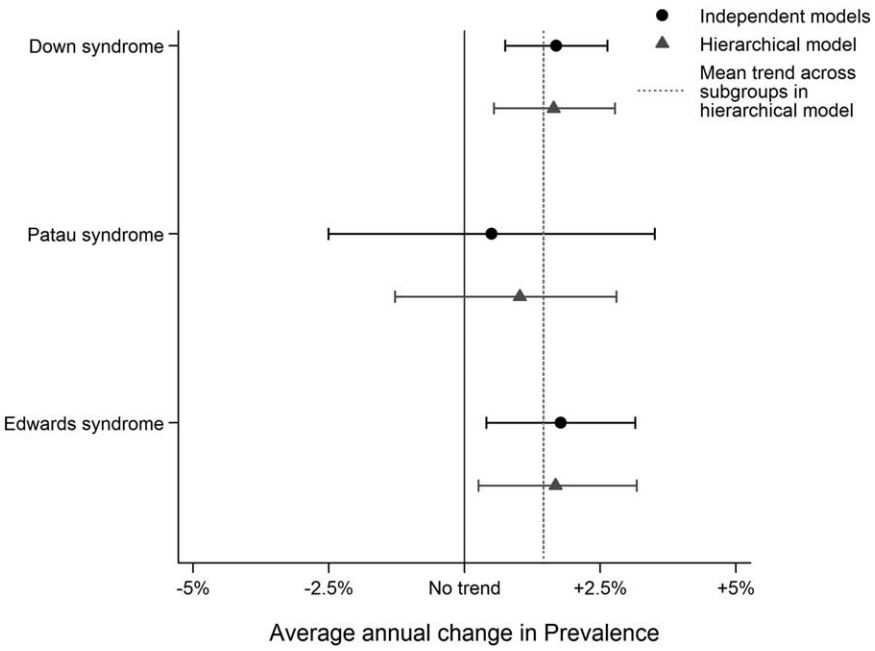


FIGURE 2. *Estimated average annual trends in prevalence of chromosomal anomaly subgroups with 95% posterior credibility intervals.*

Discussion

For all examples of congenital anomaly subgroups considered in these analyses, estimated trends in prevalence were similar whether considering anomalies separately (individual models) or together (hierarchical model). Identified trends were consistent with other studies. Increasing trends in chromosomal anomalies were observed, which are known to be due to maternal age and changes in prenatal screening practices (Cocchi et al., 2010; Loane et al., 2013). NTD prevalence remained stable in EUROCAT registries, as has been observed elsewhere (Botto et al., 2006; Khoshnood et al., 2015). This might be explained by the lack of folic acid fortification in Europe and poor uptake of folic acid supplementation; in the United Kingdom, for example, under 30% of women took folic acid before their pregnancy in 2011 to 2012 (Bestwick et al., 2014).

Prevalence in three of the digestive system anomaly subgroups was found to be significantly increasing in the latest EUROCAT statistical monitoring report (EUROCAT, 2015). A similar estimated increase in prevalence in these three subgroups was observed here, although these trends did not reach statistical significance in independent models due to the smaller number of registries included. Increases in the prevalence of the CHDs single ventricle (severity group I), ToF and atrioventricular septal defect (severity group II) were consistent with previous findings (EUROCAT, 2015). Estimated decreases in prevalence of ASD and PVS, however, were not consistent with those observed in other studies, where either no significant changes or increasing trends have been observed (van der Linde et al., 2011; EUROCAT, 2015). Published prevalence estimates in CHDs are known to vary substantially due to differing definitions of cases across studies, and it is likely that the differences in estimated trends here reflect changes in reporting for these anomalies (in recent years EUROCAT have focused on only reporting ASD cases that have been confirmed after 6 months of age) or differing prenatal screening practices in this particular set of registries (Hoffman and Kaplan, 2002; Garne et al., 2012; Baardman et al., 2014).

Hierarchical models have proven useful in the field of pharmacovigilance, where they have been used in the detection of potential adverse drug reactions (Berry and Berry, 2004; Xia et al., 2011; Crooks et al., 2012). Natural hierarchies in drug and adverse event coding have been used to group similar drugs or adverse events together, such that models for each drug-adverse event combination incorporate information from analyses of other similar drugs and adverse events (Prieto-Merino et al., 2011). In this study, the same rationale was applied to congenital anomalies; however, the situation here was different compared with that for adverse drug reactions, where the hierarchical classification systems may provide more natural hierarchies than the grouping of anomalies according to the defined subgroups.

Indeed, the EUROCAT subgroup coding hierarchy provides sets of anomalies that are too heterogeneous in practice to be grouped together when analyzing changes in prevalence. This is because the "shrinkage," a key feature of hierarchical models (Gelman and Hill, 2007) whereby the estimated trend for each subgroup is influenced toward the average trend over all subgroups in the model, will largely pull estimates toward the null if there is a mixture of increasing and decreasing trends, as was the case

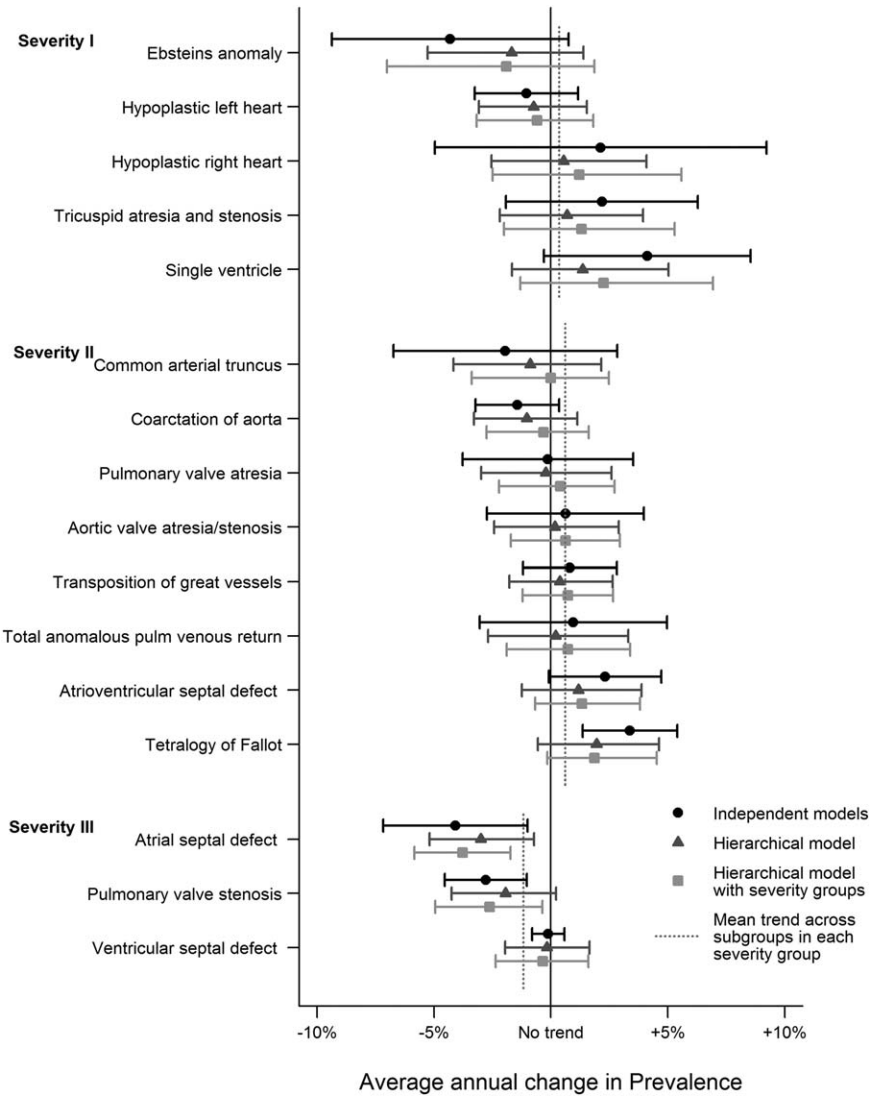


FIGURE 3. *Estimated average annual trends in prevalence of congenital heart defect subgroups with 95% posterior credibility intervals.*

for CHD and digestive anomalies. It is possible, therefore, that potential changes in prevalence in analyses of groups of anomalies such as these could actually be masked by hierarchical models.

On the other hand, this shrinkage can help control estimates based on small counts by including information from the rest of the group. Moreover, this can be thought of as a natural "penalization" considering that a hierarchical model is simultaneously looking for changes in prevalence in numerous subgroups, compared with individual models where this multiple testing aspect is not taken into account (and several false positive results are, therefore, likely). For a group where the mean trend across subgroups is close to the null, this penalization will mean that the estimated trend is no longer a "signal" in the hierarchical model, for example as seen for the CHD ToF in severity group II (Fig. 3). For a group where the mean trend is not so close to the null, however, this penalization might actually lead to an increase in the strength and precision of a signal, for example for ASD in severity group III (Fig. 3). Furthermore, the same signal might be reduced or enhanced depending on which grouping is used; for example, the trend in PVS is attenuated if considering all CHD groups together but maintained if also including the severity grouping in the hierarchical model (Fig. 3). This highlights how the posterior distribution is sensitive to the prior information, which here is the way the groups have been defined.

EUROCAT subgroups that were considered to be related, for example etiologically similar or in the same organ system class, were still found to vary considerably in terms of their differing proportional yearly changes in prevalence. It is well known (Greenland, 2000; Gelman and Hill, 2007), and has been seen here in the case of NTDs and autosomal trisomy groups of anomalies, that a random effects model with

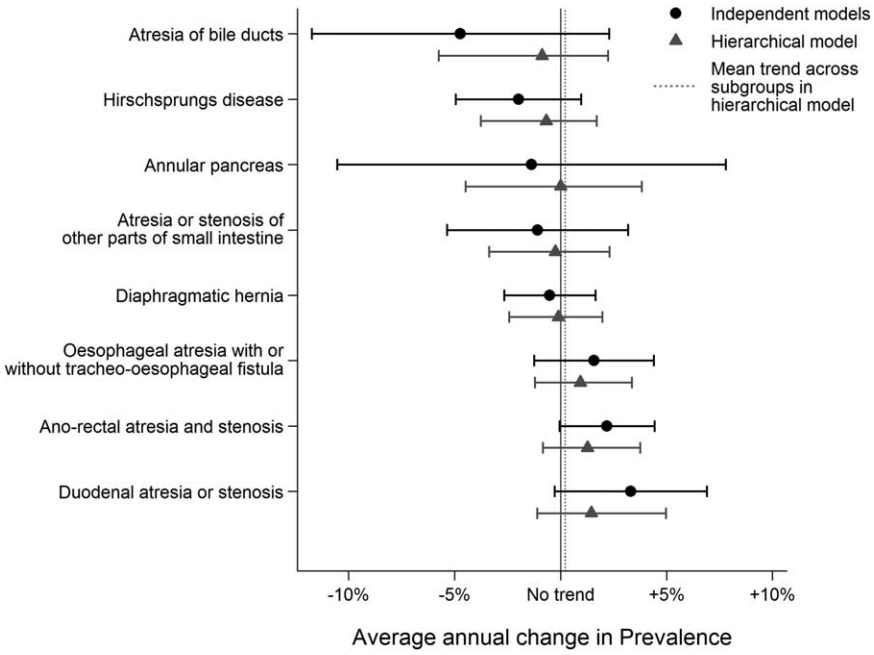


FIGURE 4. *Estimated average annual trends in prevalence of digestive system subgroups with 95% posterior credibility intervals.*

only three levels for the random effect parameter does not perform well, with such models showing poor convergence and over-parameterization. There may be other larger groups of anomalies that are similar enough to be analyzed together that were not considered here, and in fact there are known relationships between anomalies that lie within different groups of the EUROCAT hierarchy. In addition to NTDs, for example, there are several other anomalies across different body systems that are known to be sensitive to folate levels during pregnancy, including CHDs, clefts, and limb reduction defects (Wilson et al., 2015). If there was evidence that folate levels had been increasing in Europe, then it would have been useful to have analyzed all these anomalies as a hierarchical model. However, from examining the NTDs alone here and in other studies, no such change has occurred in Europe; hence, such models were not further investigated. Similarly, EUROCAT now includes a VATER/VACTERL association subgroup that comprises anomalies of the vertebra, anal atresia, CHDs, trachea-esophageal fistula, esophageal atresia, radial anomaly, and limb defect, which are known to occur together more frequently than expected by chance. However, the heterogeneity of trends in just the CHD component of this subgroup indicates that hierarchical models are not likely to add any useful information to such an analysis.

When examining congenital anomaly prevalence, there are many factors that are likely to have an influence, such as reporting, case ascertainment, or screening practices. Hierarchical models might be more relevant, then, when considering the risks of specific exposures in relation to prevalence of congenital anomalies. It would, therefore, be worthwhile investigating the application of hierarchical models in such situations, for example, when looking at the risk of medications taken during the first trimester of pregnancy.

STRENGTHS AND LIMITATIONS OF THIS STUDY

EUROCAT registries collect data that is ascertained from multiple sources and includes information on all major structural congenital and chromosomal anomalies (Boyd et al., 2011; Loane et al., 2011a), providing high quality population-based data across multiple European countries and allowing the inclusion of a large number of congenital anomaly cases covering over four million births over 10 years for this study. EUROCAT registries include information on cases of prenatal diagnosis followed by termination of pregnancy, enabling the inclusion of cases that would otherwise have gone undiagnosed, or unreported among spontaneous abortions or stillbirths. One potential limitation of this study is that it was not possible to include data from all of the EUROCAT member registries in these analyses; hence, some trends that were seen in the latest statistical monitoring report did not reach statistical significance here, likely due to the smaller sample sizes included. However, it does not seem likely that increasing the sample size would have improved the performance of hierarchical models.

CONCLUSIONS

In summary, the hierarchical models considered here demonstrated that sharing information between subgroups of anomalies can provide a sensible "penalization," helping to avoid false positive signals by shrinking the estimated trends toward the null when there is no evidence of other trends in the rest of the group, while maintaining signals of changes in

prevalence when there are others in the group. Using the EUROCAT hierarchy of anomaly subgroups, however, presented no substantial differences between the independent analyses of each subgroup and hierarchical models. When using EUROCAT subgroups for analysis, therefore, considering each congenital anomaly separately remains an appropriate method for the detection of potential changes in prevalence by relevant surveillance systems. Hierarchical models do, however, remain an interesting and potentially useful alternative method of analysis when considering the risks of specific exposures in relation to the prevalence of congenital anomalies, and this could be investigated in other studies.

Acknowledgments

EUROCAT registries are funded as described in Paper 6 of Report 9 "EUROCAT Member Registries: Organization and Activities" (http://onlinelibrary.wiley.com/doi/10.1002/bdra.20775/pdf). We thank the people throughout Europe involved in providing and processing information, including affected families, clinicians, health professionals, medical record clerks, and registry staff.

References

Baardman ME, du Marchie Sarvaas GJ, de Walle HE, et al. 2014. Impact of introduction of 20-week ultrasound scan on prevalence and fetal and neonatal outcomes in cases of selected severe congenital heart defects in The Netherlands. Ultrasound Obstet Gynecol 44:58–63.

Berry SM, Berry DA. 2004. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics 60:418–426.

Bestwick JP, Huttly WJ, Morris JK, Wald NJ. 2014. Prevention of neural tube defects: a cross-sectional study of the uptake of folic acid supplementation in nearly half a million women. PLoS One 9:e89354.

Botto LD, Lisi A, Bower C, et al. 2006. Trends of selected malformations in relation to folic acid recommendations and fortification: an international assessment. Birth Defects Res A Clin Mol Teratol 76:693–705.

Boyd PA, Haeusler M, Barisic I, et al. 2011. Paper 1: the EUROCAT network--organization and processes. Birth Defects Res A Clin Mol Teratol 91(Suppl 1):S2–S15.

Casella G, George EI. 1992. Explaining the Gibbs Sampler. Am Stat 46:167–174.

Cocchi G, Gualdi S, Bower C, et al. 2010. International trends of Down syndrome 1993-2004: births in relation to maternal age and terminations of pregnancies. Birth Defects Res A Clin Mol Teratol 88:474–479.

Crooks CJ, Prieto-Merino D, Evans SJ. 2012. Identifying adverse events of vaccines using a Bayesian method of medically guided information sharing. Drug Saf 35:61–78.

Dolk H. 2005. EUROCAT: 25 years of European surveillance of congenital anomalies. Arch Dis Child Fetal Neonatal Ed 90:F355–F358.

EUROCAT. 2005. EUROCAT Guide 1.3: instructions for the registration and surveillance of congenital anomalies. Northeren Ireland, United Kingdom: EUROCAT Central Registry, University of Ulster.

EUROCAT. 2009. Special report, congenital heart defects in Europe: 2000-2005. ISPRA (VA): Italy: EUROCAT.

EUROCAT. 2015. EUROCAT Statistical Monitoring Report 2012. [ONLINE] Available at: http://www.eurocat-network.eu/content/Stat-Mon-Report-2012.pdf. Accessed February 15, 2016.

EUROCAT. 2016. What is EUROCAT? [ONLINE] Available at: http://www.eurocat-network.eu/aboutus/whatiseurocat/whatiseurocat. Accessed February 23, 2016.

Garne E, Olsen MS, Johnsen SP, et al. 2012. How do we define congenital heart defects for scientific studies? Congenit Heart Dis 7:46–49.

Gelman A, Hill J. 2007. Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press.

Greenland S. 2000. Principles of multilevel modelling. Int J Epidemiol 29:158–167.

Greenlees R, Neville A, Addor MC, et al. 2011. Paper 6: EUROCAT member registries: organization and activities. Birth Defects Res A Clin Mol Teratol 91(Suppl 1):S51–S100.

Hoffman JI, Kaplan S. 2002. The incidence of congenital heart disease. J Am Coll Cardiol 39:1890–1900.

Khoshnood B, Loane M, Walle H, et al. 2015. Long term trends in prevalence of neural tube defects in Europe: population based study. BMJ 351:h5949.

Khoury MJ, Botto L, Mastroiacovo P, et al. 1994. Monitoring for multiple congenital anomalies: an international perspective. Epidemiol Rev 16:335–350.

Loane M, Dolk H, Garne E, Greenlees R. 2011a. Paper 3: EUROCAT data quality indicators for population-based registries of congenital anomalies. Birth Defects Res A Clin Mol Teratol 91(Suppl 1):S23–S30.

Loane M, Dolk H, Kelly A, et al. 2011b. Paper 4: EUROCAT statistical monitoring: identification and investigation of ten year trends of congenital anomalies in Europe. Birth Defects Res A Clin Mol Teratol 91(Suppl 1):S31–S43.

Loane M, Morris JK, Addor MC, et al. 2013. Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: impact of maternal age and prenatal screening. Eur J Hum Genet 21:27–33.

McBride WG. 1961. Thalidomide and congenital abnormalities. Lancet 278:1358.

Plummer M. 2003. JAGS: a program for analysis of bayesian graphical models using Gibbs sampling; 2003 March 20–22.

Plummer M, Best N, Cowles K, Vines K. 2003. CODA: convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Prieto-Merino D, Quartey G, Wang J, Kim J. 2011. Why a Bayesian approach to safety analysis in pharmacovigilance is important. Pharm Stat 10:554–559.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

van der Linde D, Konings EE, Slager MA, et al. 2011. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol 58:2241–2247.

Wilson RD, Audibert F, Brock JA, Carroll J, Cartier L, Gagnon A, Johnson JA, Langlois S, Murphy-Kaulbeck L, Okun N, Pastuck M, Deb-Rinker P, Dodds L, Leon JA, Lowel HL, Luo W, MacFarlane A, McMillan R, Moore A, Mundle W, O'Connor D, Ray J, Van den Hof M. 2015. Preconception Folic Acid and Multivitamin Supplementation for the Primary and Secondary Prevention of Neural Tube Defects and Other Folic Acid-Sensitive Congenital Anomalies. SOGC Clinical Practice Guideline no. 324, May 2015. J Obstet Gynaecol Can 37(6):534–552.

World Health Organization. 2014. Fact sheet N° 370 Congenital anomalies. [ONLINE] Available at: http://www.who.int/mediacentre/factsheets/fs370/en/. Accessed January 28, 2016.

Xia HA, Ma H, Carlin BP. 2011. Bayesian hierarchical modeling for detecting safety signals in clinical trials. J Biopharm Stat 21:1006–1029.

Appendix B: Supplementary material for Chapter 5

B1. Chapter 5 figures with ATC4 groupings

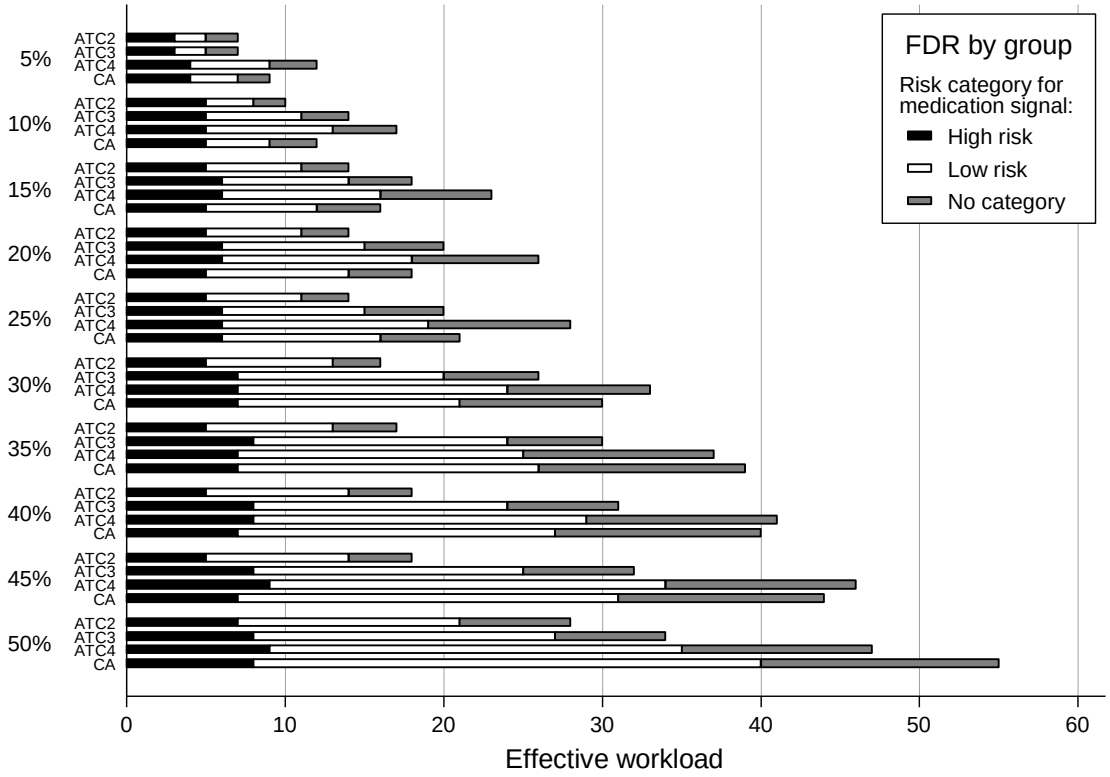


Figure B1. Effective workload and the number of medication signals in each risk category using the group BH procedure. Results are for grouping of medication-CA combinations by ATC2, ATC3 codes and CA subgroups according to cut-offs for FDR level from 5% to 50% (Figure 5.2 with ATC4 groupings added).

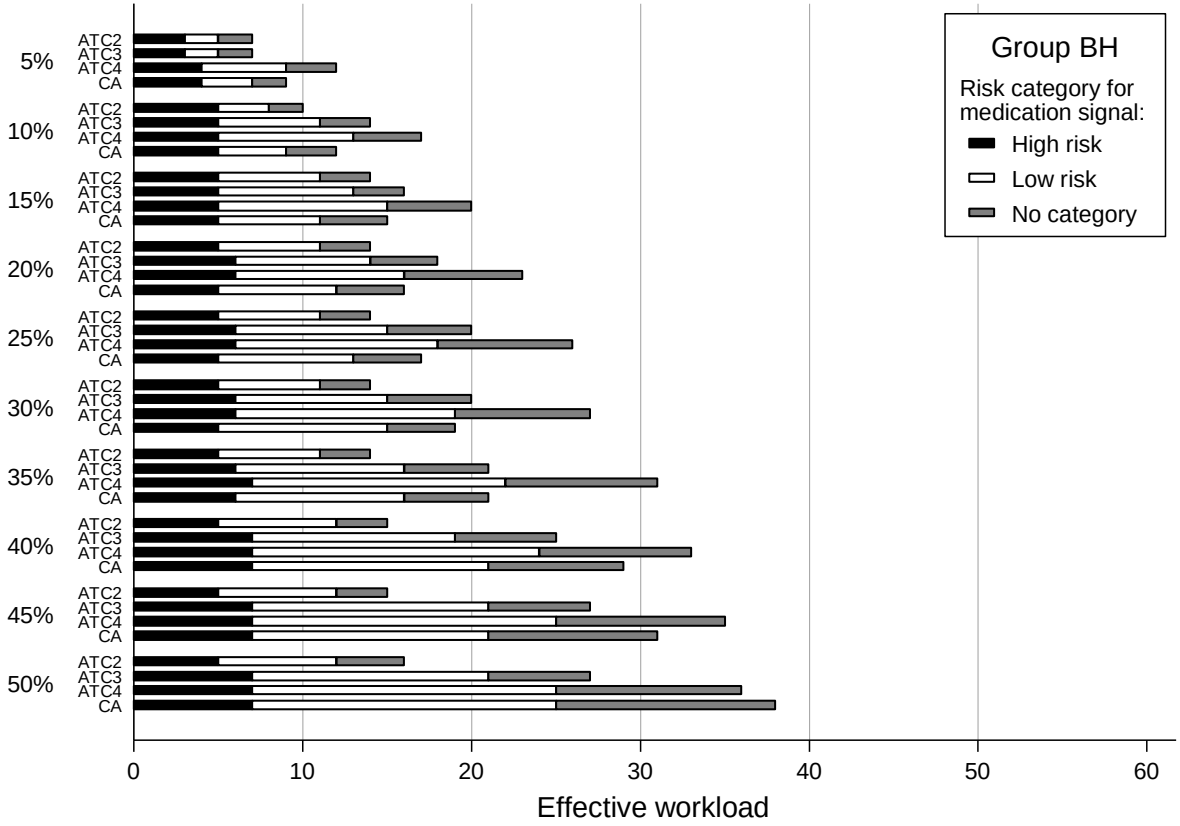


Figure B2. Effective workload and the number of medication signals in each risk category using the group BH procedure. Results are for grouping of medication-CA combinations by ATC2, ATC3, ATC4 codes and CA subgroups according to cut-offs for FDR level from 5% to 50% (Figure 5.3 with ATC4 groupings added).

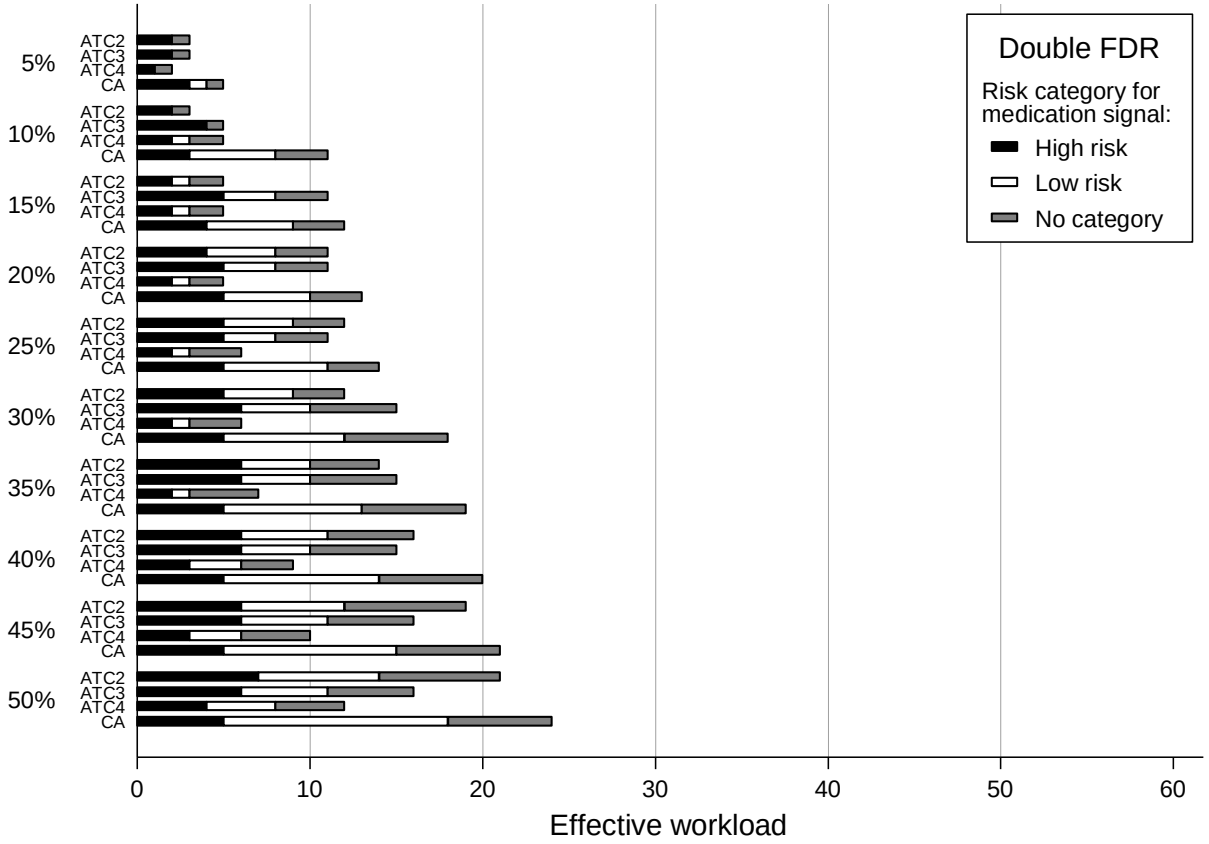


Figure B3. Effective workload and the number of medication signals in each risk category using the double FDR procedure. Results are for grouping of medication-CA combinations by ATC2, ATC3, ATC4 codes and CA subgroups according to cut-offs for FDR level from 5% to 50% (Figure 5.4 with ATC4 groupings added).

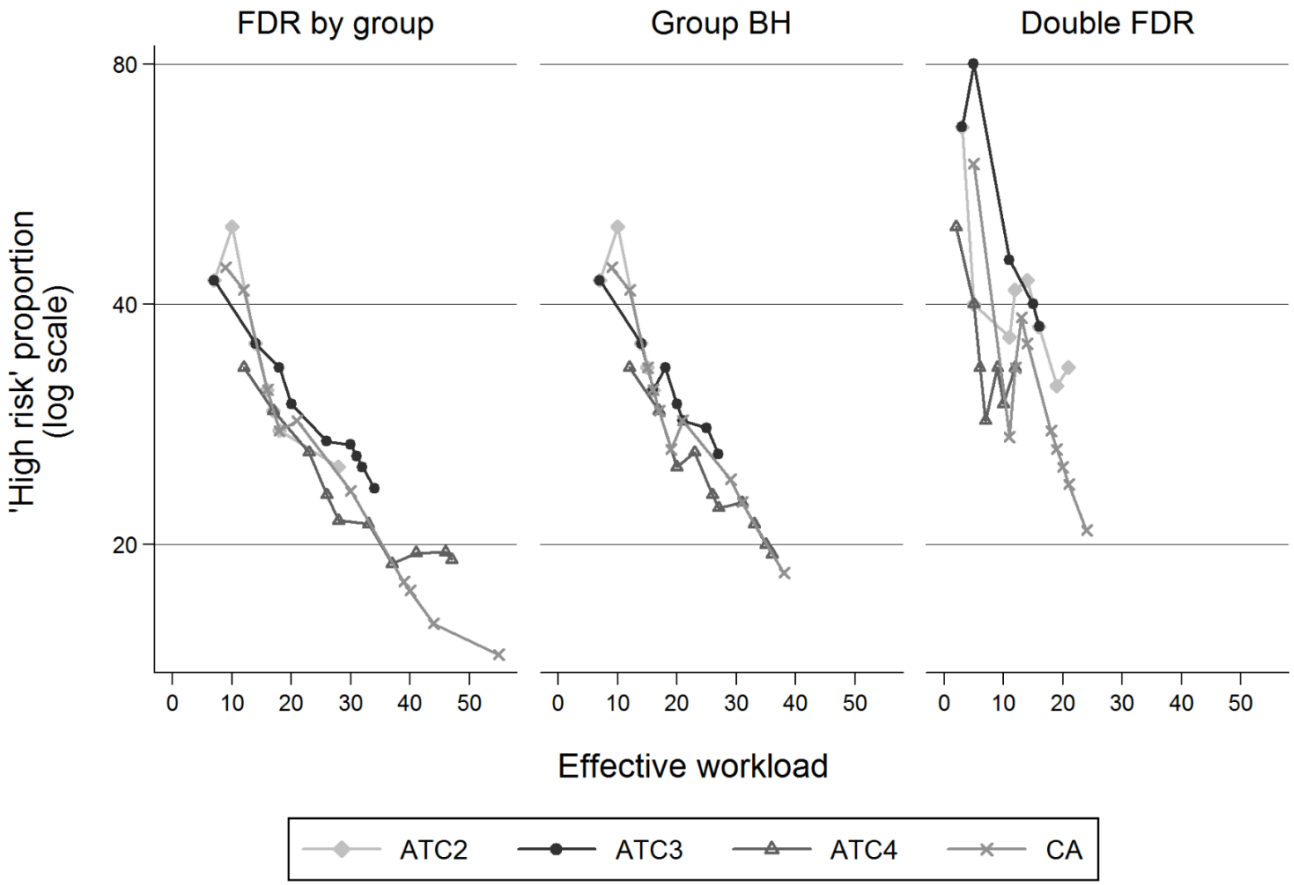


Figure B4. "High risk" proportion (percent of all signals that are in the "high risk" category) against effective workload (total number of signals) for FDR by group, group BH and double FDR methods. Grouping is by ATC2, ATC3, ATC4 codes and by CA subgroup, and each point corresponds to a different level of FDR for that type of grouping in 5% increments from 5% to 50% (Figure 5.5 with ATC4 groupings added).

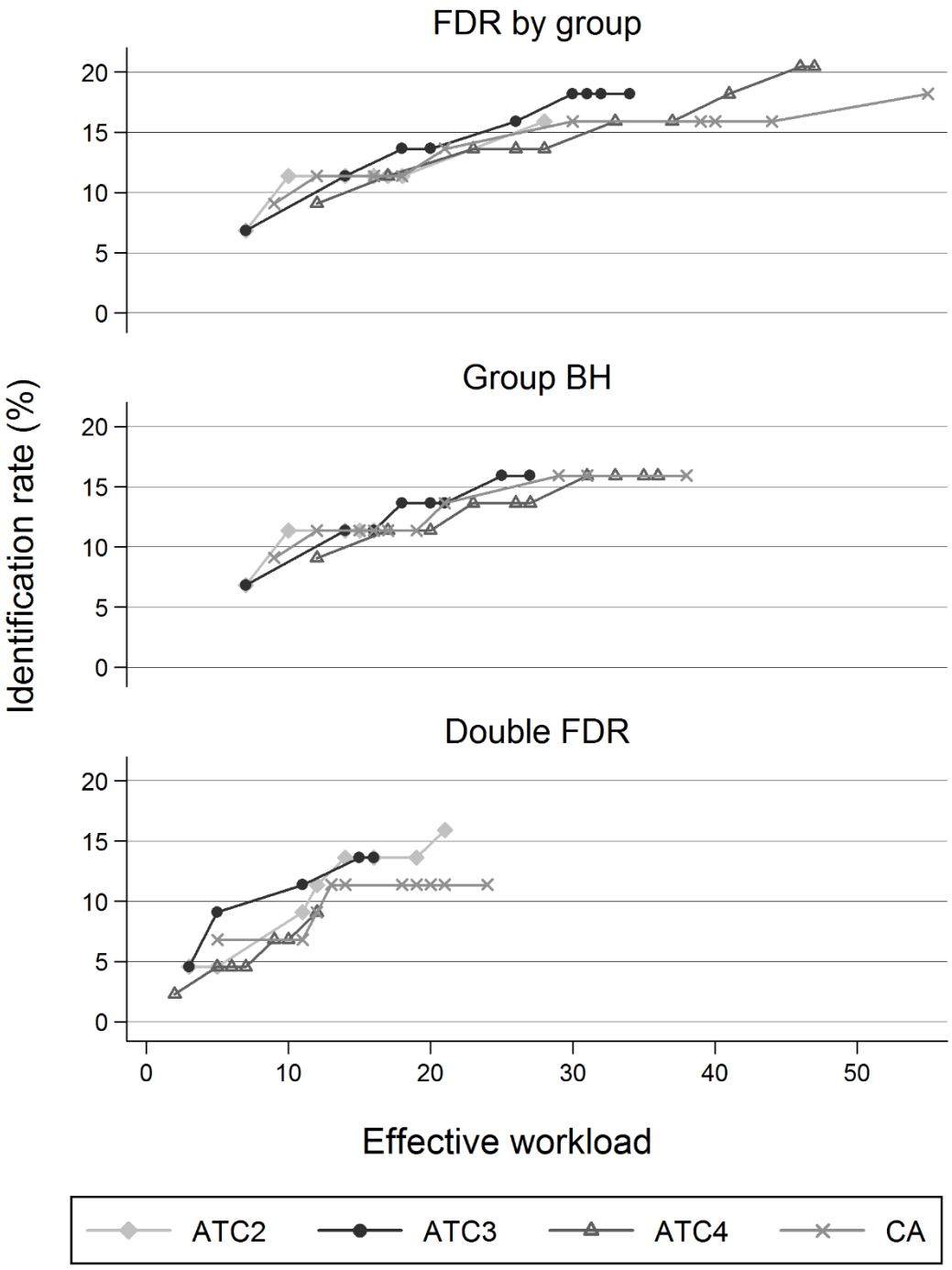


Figure B5. Identification rate (proportion of all "high risk" combinations identified as signals) against effective workload (the total number of signals) for FDR by group, group BH and double FDR methods. Grouping is by ATC2, ATC3 codes and by CA subgroup, and each point corresponds to a different level of FDR for that type of grouping in 5% increments from 5% to 50% (Figure 5.6 with ATC4 groupings added).

B2. Stata code used to run a double FDR procedure with ATC3 groupings

```
*-----------------------------------------------------------*

** RUNNING Fisher's exact tests for each medication-CA combination

clear
set obs 1
gen CA str1 = "a"
save tmp,replace

use "ALL_meds_anoms_long.dta", clear
gen v1=1

foreach i of varlist A01AA01-V07AB99 {
foreach k of varlist a1-a36 {

preserve

 *excluding data for registries with no records of medication i
  egen Exp = max(`i'), by(centre)
  drop if Exp == 0
  drop Exp
  di "drug: `i', CA: `k'"

 *excluding data for registries with no records of the CA k
  egen Exp = max(`k'), by(centre)
  drop if Exp == 0 & r(N) != 0
  drop Exp
  count

 *Fisher's exact test
  capture cs `k' `i', exact

 *storing the estimates
  gen p = `r(p1_exact)'
  gen rr str30 = string(round(`r(rr)', 0.01)) + " (" +
string(round(`r(lb_rr)', 0.01)) + " - " + string(round(`r(ub_rr)', 0.01))
+ ")"
  loc lab`k': variable label `k'
  gen CA = "`lab`k''"
  collapse (sum) v1 `k', by (`i' p rr CA)
  rename (v1 `k') (control case)
  reshape wide control case, i(rr p CA) j(`i')
  replace control0 = control0 - case0
  replace control1 = control1 - case1
  rename (control0 control1 case0 case1) (AnoDno AnoDyes AyesDno AyesDyes)
  gen ATC = "`i'"
  append using tmp
  save tmp, replace
restore
}
}
*-----------------------------------------------------------*
```

```
*some data sorting
use tmp, clear
drop if CA == "a"
drop if AyesDno == 0 & AyesDyes == 0
save InterimResATC5, replace

*-----------------------------------------------------------*
*The following do file flags medication-CA combinations that are signals
using
      -  no adjustment for multiple testing
      -  single FDR adjustment
      -  double FDR adjustment

This is done in a loop for a range of FDR cut-off values (here 5%, 10%,
20%, 30%, 40% and 50%)

*------------------- START OF LOOP ----------------------*

local siglevel "05 1 2 3 4 5"
foreach i of local siglevel {
local alpha=0.`i'
count
local tot=r(N)

*----------------------------------*

*combinations with an unadjusted P-value < α%
gen f_noadj`i'=1 if p<`alpha'
tab f_noadj`i'

*single FDR (one-step BH procedure)
qqvalue p, method(simes) qvalue(pbh`i')
gen f_bh`i'=1 if pbh`i'<`alpha'
tab f_bh`i'

*----------------------------------*

***double FDR, STEP 1: define a representative P-value for each group
pistar by choosing the smallest FDR-adjusted P-value within each group

*FDR adjustment within each group
bys grp: qqvalue p, method(simes) qvalue(pgrp`i')

*tag the smallest FDR-adjusted P-value in each group
bys grp: egen pistar`i'=min(pgrp`i')
replace pistar`i'=. if tag!=1

*apply an FDR adjustment over the representative P-values from each group
qqvalue pistar`i', method(simes) qvalue(pistt`i')

*signals are flagged if the representative P-value is < alpha AND the P-
value from step 2 is < alpha
gen tmp=1 if pistt`i'<`alpha'
bys grp:egen F`i'=max(tmp)
drop tmp
sort grp ATC CA
```

```
***double FDR, STEP 2: applying a single FDR adjustment to the P-values in
F
qqvalue p if F`i'==1, method(simes) qvalue(pdfdr`i')

gen f_dfdr`i'=1 if pdfdr`i'<`alpha'
di 100*(r(N))/`tot'

}

*---------------------- END OF LOOP -------------------------*
```

Appendix C: Supplementary material for Chapter 6

C1. Code used to specify BHMs in JAGS and R for chapter 6

No information sharing: Poisson model

```
for (i in 1:dmax) {
  for (j in 1:CAmax){

    c[i,j] ~ dpois(p[i,j])
    p[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (CA[j]*d[i])/N
    PRR[i,j] <- exp(lambda[i,j])
    lambda[i,j] ~ dnorm(0, 0.33)
  }
}
```

No information sharing: Negative Binomial model

```
for (i in 1:dmax) {
  for (j in 1:CAmax){

    c[i,j] ~ dnegbin(p[i,j], r[i,j])
    p[i,j] <- r[i,j] / (r[i,j] + mu[i,j])
    mu[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (CA[j]*d[i])/N
    PRR[i,j] <- exp(lambda[i,j])
    lambda [i,j] ~ dnorm(0, 0.33)
    r[i,j] ~ dunif(0,1000)
  }
}
```

Information sharing for medications only: Poisson model

```
for (i in 1:dmax) {
  for (j in 1:CAmax){

    c[i,j] ~ dpois(p[i,j])
    p[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (d[i]*CA[j])/N
    PRR[i,j] <- exp(lambda[i,j])

    #grouping by ATC medication codes
    lambda[i,j] ~ dnorm(theta[groupd[i]], tau[groupd[i]])
  }
}

for (k in 1:groupdmax){
    theta[k] ~ dnorm(0, 0.33)
    tau[k] <- 1/sigma2[k]
    sigma2[k] <- pow(sigma[k],2)
    sigma[k] ~ dunif(0,100)
}
```

Information sharing for medications only: Negative binomial model

```
for (i in 1:dmax) {
 for (j in 1:CAmax){
    c[i,j] ~ dnegbin(p[i,j], r)
    p[i,j] <- r / (r + mu[i,j])
    mu[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (CA[j]*d[i])/N
    PRR[i,j] <- exp(lambda[i,j])
    lambda[i,j] ~ dnorm(theta[groupd[i]], tau[groupd[i]])
  }
}
 r ~ dunif(0, 1000)

for (k in 1:groupdmax){
    theta[k] ~ dnorm(0, 0.33)
    tau[k] <- 1/sigma2[k]
    sigma2[k] <- pow(sigma[k],2)
    sigma[k] ~ dunif(0,100)
}
```

Information sharing for CAs only: Poisson model

```
for (i in 1:dmax) {
  for (j in 1:CAmax){

    c[i,j] ~ dpois(p[i,j])
    p[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (CA[j]*d[i])/N
    PRR[i,j] <- exp(lambda[i,j])

    #grouping by type of CA
    lambda[i,j] ~ dnorm(theta[groupCA[j]], tau[groupCA [j]])
  }
}

for (l in 1:groupCAmax){
    theta[l] ~ dnorm(0, 0.33)
    tau[l] <- 1/sigma2[l]
    sigma2[l] <- pow(sigma[l],2)
    sigma[l] ~ dunif(0,100)
}
```

Information sharing for CAs only: Negative binomial

```
for (i in 1:dmax) {
 for (j in 1:CAmax){
    c[i,j] ~ dnegbin(p[i,j], r)
    p[i,j] <- r / (r + mu[i,j])
    mu[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (CA[j]*d[i])/N
    PRR[i,j] <- exp(lambda[i,j])
   lambda[i,j] ~ dnorm(theta[groupCA[j]], tau[groupCA[j]])
  }
}
 r ~ dunif(0, 1000)
```

```
for (l in 1:groupCAmax){
    theta[l] ~ dnorm(0, 0.33)
    tau[l] <- 1/sigma2[l]
    sigma2[l] <- pow(sigma[l],2)
    sigma[l] ~ dunif(0,100)
}
```

Information sharing in two dimensions (medications and CAs): Poisson model

```
for (i in 1:dmax) {
  for (j in 1:CAmax){
    c[i,j] ~ dpois(p[i,j])
    p[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (d[i]*CA[j])/N
    PRR[i,j] <- exp(lambda[i,j])

   #grouping by type of CA and ATC3 medication codes
   lambda[i,j] ~ dnorm(theta[groupd[i],groupCA[j]],
tau[groupd[i],groupCA[j]])
  }
}

for (k in 1:groupdmax){
  for (l in 1:groupCAmax){
    theta[k,l] ~ dnorm(0, 0.33)
    tau[k,l] <- 1/sigma2[k,l]
    sigma2[k,l] <- pow(sigma[k,l],2)
    sigma[k,l] ~ dunif(0,100)
  }
}
```

Information sharing in two dimensions (medications and CAs): Negative Binomial model

```
for (i in 1:dmax) {
 for (j in 1:CAmax){
    c[i,j] ~ dnegbin(p[i,j], r)
    p[i,j] <- r / (r + mu[i,j])
    mu[i,j] <- (PRR[i,j])*(E[i,j])
    E[i,j] <- (CA[j]*d[i])/N
    PRR[i,j] <- exp(lambda[i,j])
    lambda[i,j] ~ dnorm(theta[groupd[i],groupCA[j]],
tau[groupd[i],groupCA[j]])
  }
}
 r ~ dunif(0, 1000)

for (k in 1:groupdmax){
 for (l in 1:groupCAmax){
    theta[k,l] ~ dnorm(0, 0.33)
    tau[k,l] <- 1/sigma2[k,l]
    sigma2[k,l] <- pow(sigma[k,l],2)
    sigma[k,l] ~ dunif(0,100)
 }
}
```

C2. Sensitivity analyses for BHMs in chapter 6

Table C1. Observed and expected counts for congenital heart defects, oro-facial clefts and limb CAs in association with human insulin medication A10AC01. The marginal total count for A10AC01 is $c_{i.} = 141$ and the total count for all exposures across the dataset is $N = 26{,}765$.

CA subgroup	Marginal total for CA ($c_{.j}$)	Counts for CA in combination with A10AC01	
		Observed	Expected [a]
Congenital heart defects			
Aortic valve atresia/stenosis	197	2	1.0
Atrial septal defect	2,046	21	10.8
Atrioventricular septal defect	246	1	1.3
Coarctation of aorta	351	4	1.8
Common arterial truncus	55	1	0.3
Ebstein's anomaly	48	0	0.3
Hypoplastic right heart	28	0	0.1
Patent ductus arteriosus as only CHD in term infants	404	11	2.1
Pulmonary valve atresia	115	0	0.6
Pulmonary valve stenosis	468	4	2.5
Single ventricle	84	2	0.4
Tetralogy of Fallot	325	2	1.7
Total anomalous pulmonary venous return	48	0	0.3
Transposition of great vessels	357	6	1.9
Tricuspid atresia and stenosis	69	0	0.4
Ventricular septal defect	3,964	29	20.9
Unspecified CHDs	801	11	4.2
Oro-facial clefts			
Cleft lip with or without palate	1,047	0	5.5
Cleft palate	709	2	3.7
Limb			
Club foot - talipes equinovarus	1,293	5	6.8
Limb reduction	658	1	3.5
Polydactyly	909	3	4.8
Syndactyly	557	1	2.9

[a] *The marginal total count for A10AC01 is* $c_{i.} = 141$ *and the total count for all exposures across the dataset is* $N = 26{,}765$. *The expected count is then calculated as* $\frac{c_{.j} \, x \, 141}{26{,}765}$.

Table C2. Sensitivity of the thresholds used to define signals for Poisson BHMs in chapter 6.

Type of grouping	**PCI cut-off level**	Number of signals as displayed in Table 6.9		Signals as in Table 6.9, with **threshold raised to lower 95% PCI >2** [a]		Percentage of signals retained after change in threshold to **lower 95% PCI >2** [a]
		Combinations	Medications	Combinations	Medications	
No grouping	95%	223	159	21	21	10%
Discrete grouping by medications	95%	21	15	1	1	7%
Discrete grouping by CAs	95%	10	8	1	1	13%
Discrete grouping by medications and CAs	95%	112	71	8	8	4%

[a] *An additional restriction excluding combinations with less than 3 exposure counts changes numbers only for the Poisson BHMs with no grouping; this reduces to only 16 combinations (for 16 medications) being signals*

[b] *For the single and double FDR methods the lower 95% CI is used*

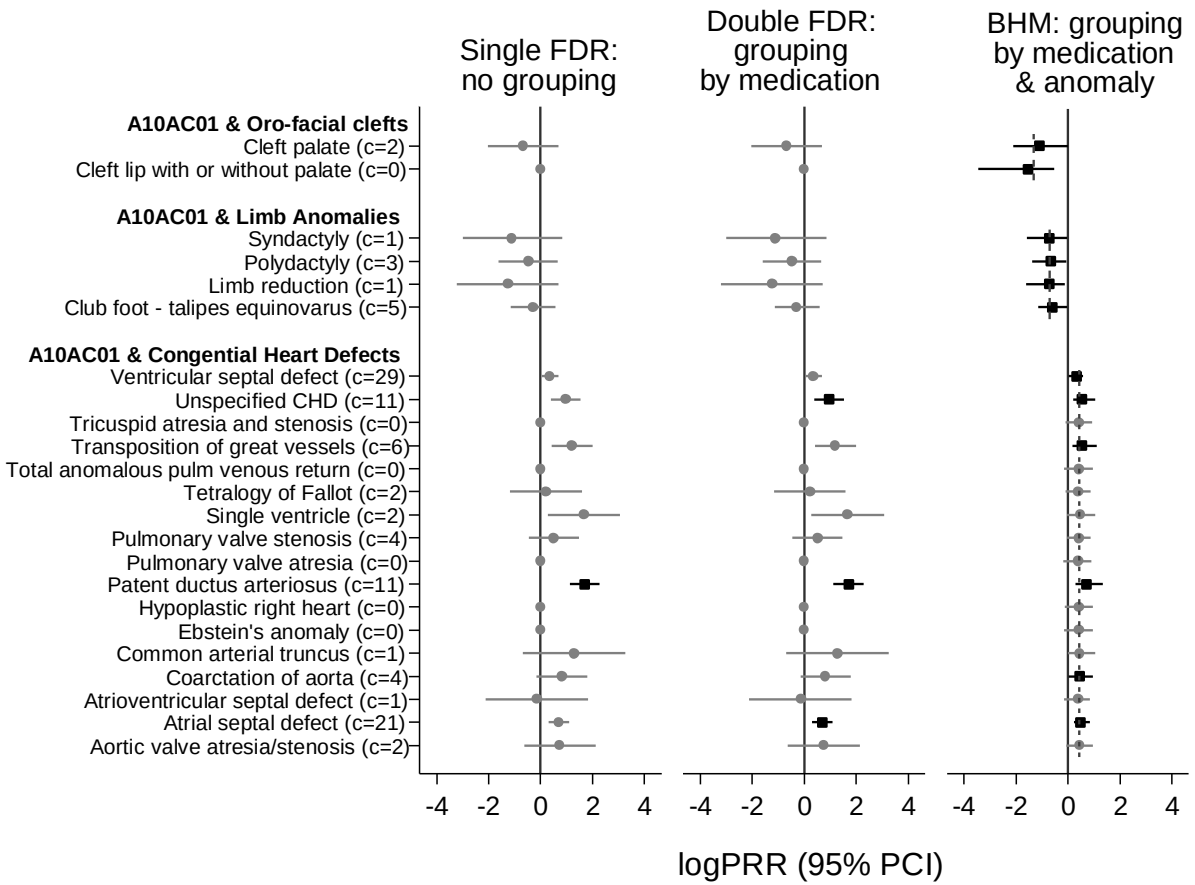


Figure C1. Example of protective associations caused by a combination of signals for the same medication in another group of CAs, and shrinkage for small groups with low cell counts: human insulin A10AC01 in combination with congenital heart defects, oro-facial clefts and limb CAs. The dashed lines in the BHM show the mean $log(PRR)$ across each group of CAs in combination with the A10A group of ATC3 medications.

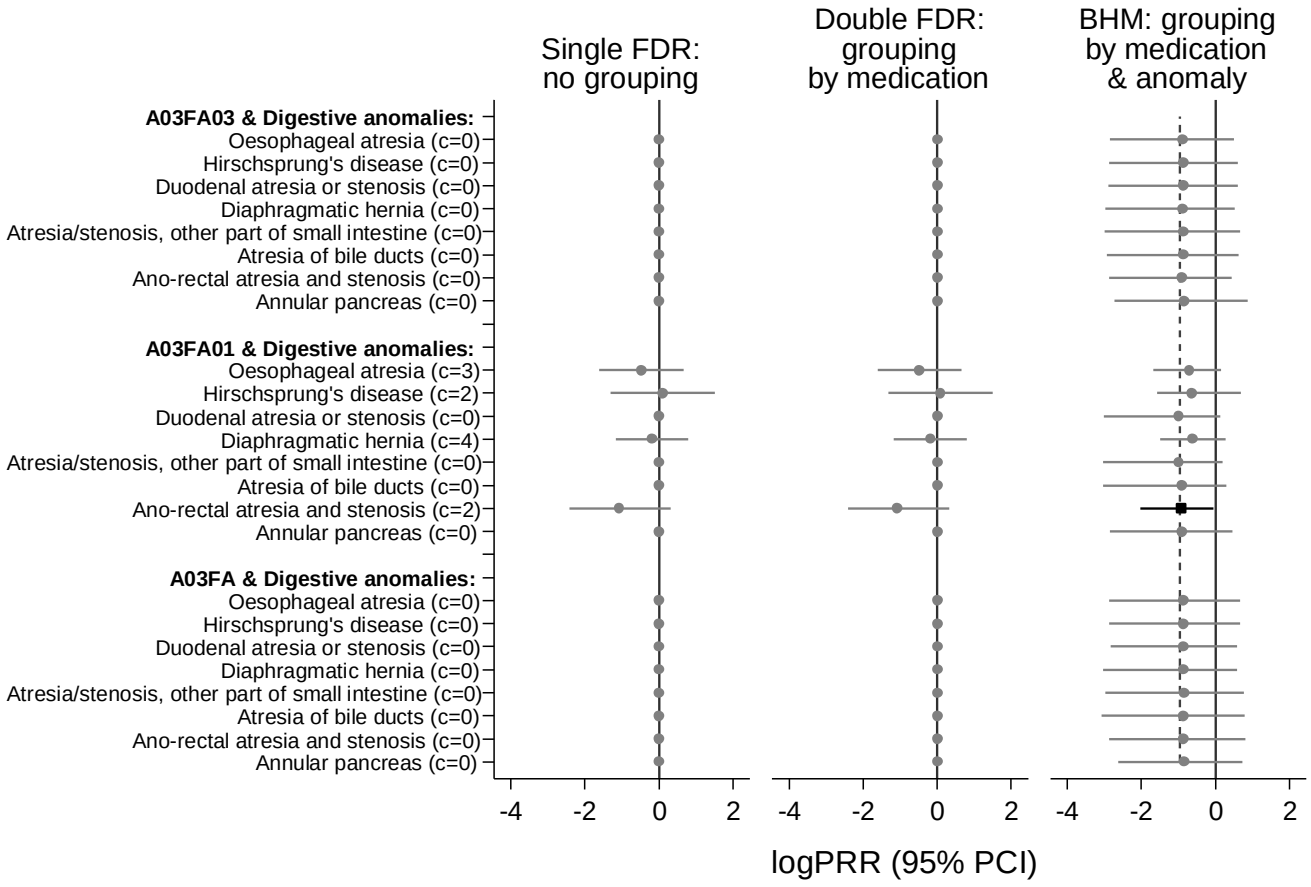


Figure C2. Example of a protective association in a medication-CA group with a high proportion of zero cell counts: eight digestive system CAs in combination with three A03F medications for functional gastrointestinal disorders. The dashed line in the BHM shows the mean $log(PRR)$ across the group of A03F medications and digestive system CAs